

# Multimodal few-shot learning & probing self-supervised speech models

LSCP, Ecole Normale Supérieure, Sep. 2023

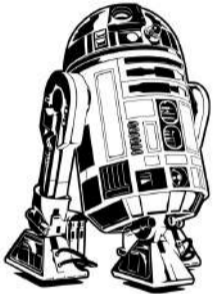
Herman Kamper

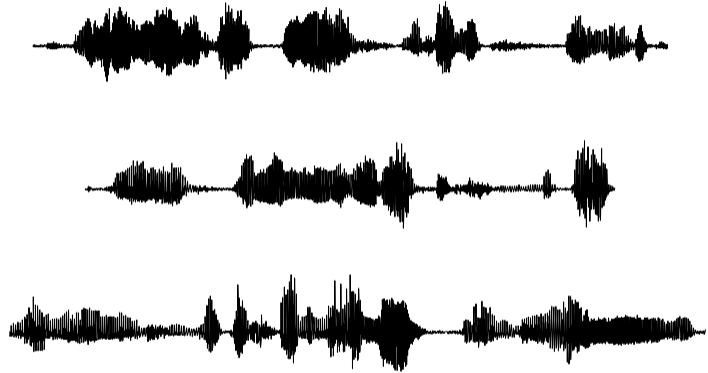
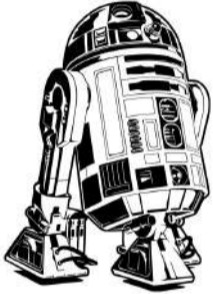
E&E Engineering, Stellenbosch University, South Africa

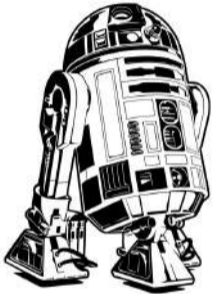
<http://www.kamperh.com/>

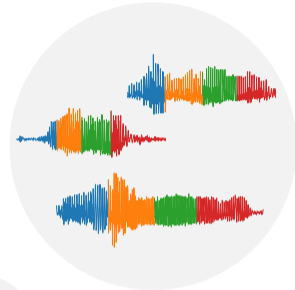
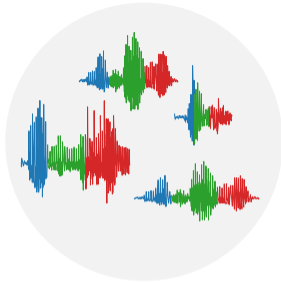
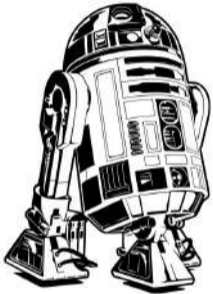
















# Why attempt to emulate language acquisition?



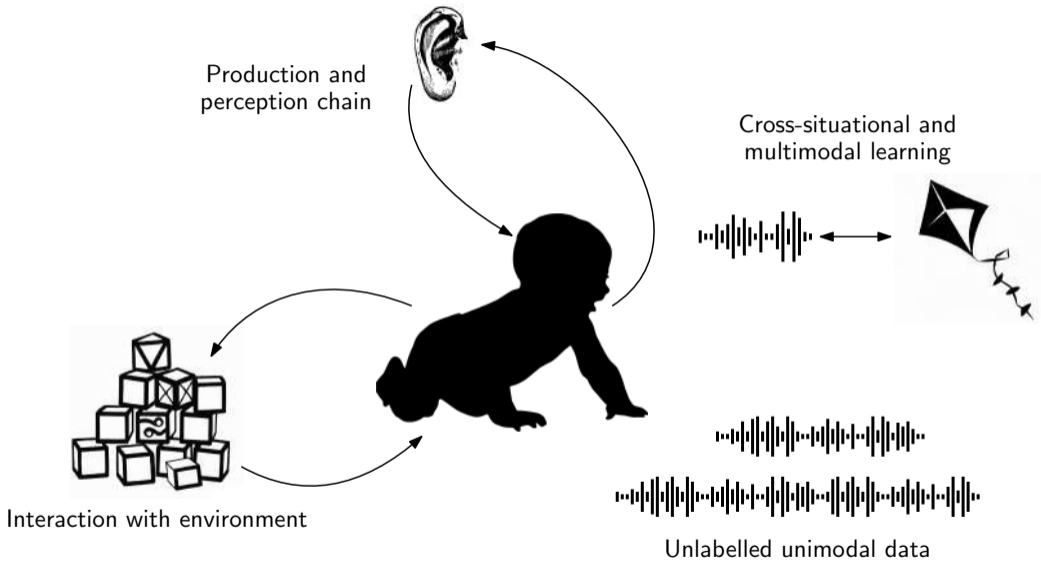
Improvements in speech technology

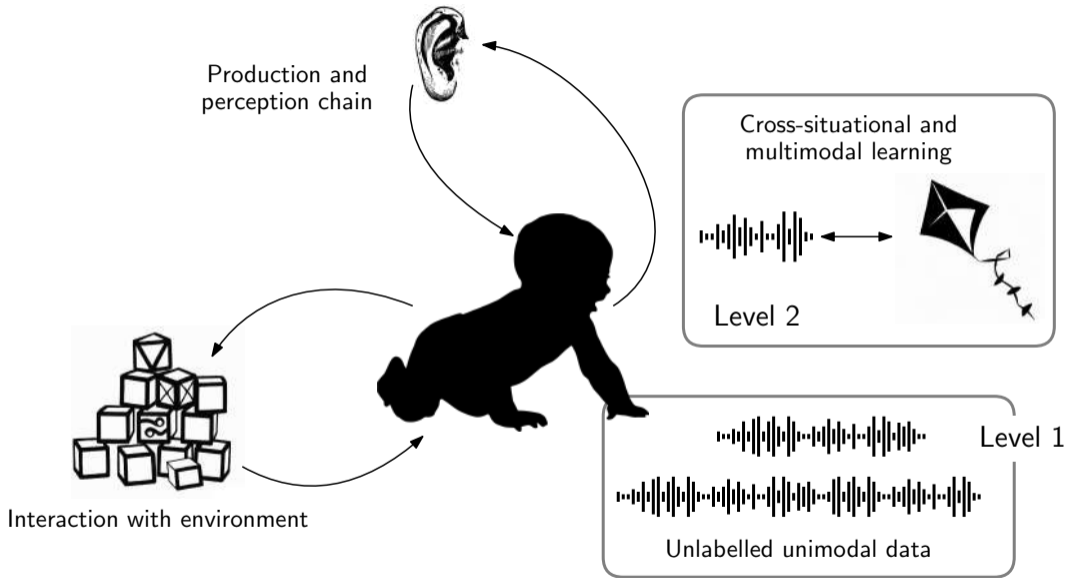


New insights and approaches for machines that learn



New insights into human learning





# 1. Multimodal few-shot learning from images and speech



Leanne  
Nortje



Kayode  
Olaleye



Dan  
Oneață

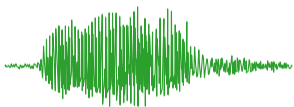
Nortje et al., "Visually grounded few-shot word acquisition with fewer shots," in *Interspeech*, 2023.

Nortje et al., "Visually grounded few-shot word learning in low-resource settings," *arXiv*, 2023.

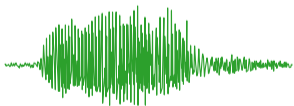


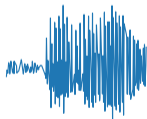
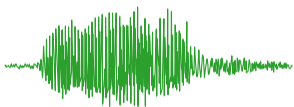






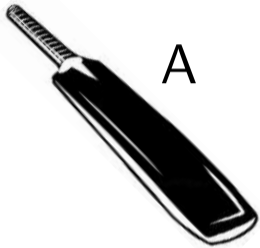












A



B

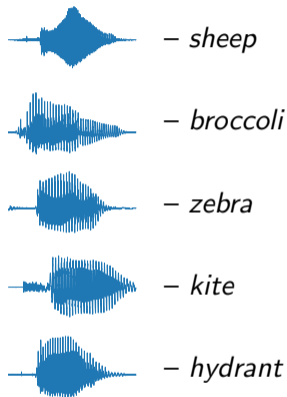


C

?



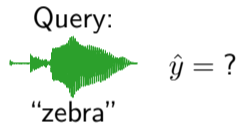
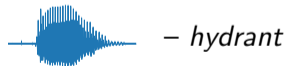
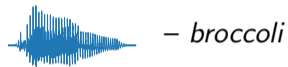
# Unimodal one-shot learning and classification



Fei-Fei et al., "One-shot learning of object categories," *TPAMI*, 2006.

Lake et al., "One-shot learning of generative speech concepts," in *CogSci*, 2014.

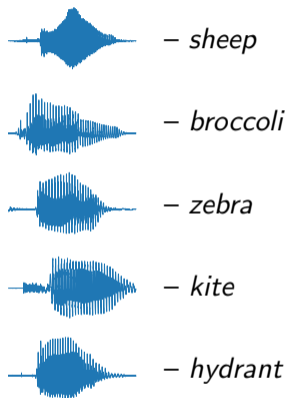
# Unimodal one-shot learning and classification



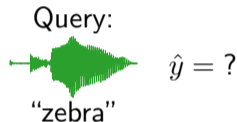
Fei-Fei et al., “One-shot learning of object categories,” *TPAMI*, 2006.

Lake et al., “One-shot learning of generative speech concepts,” in *CogSci*, 2014.

# Unimodal one-shot learning and classification



One-shot speech learning



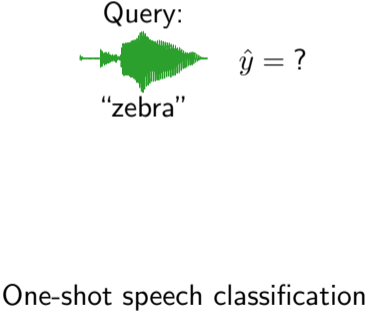
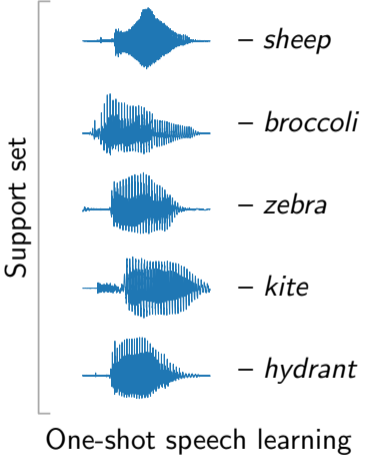
One-shot speech classification

Fei-Fei et al., "One-shot learning of object categories," *TPAMI*, 2006.

Lake et al., "One-shot learning of generative speech concepts," in *CogSci*, 2014.



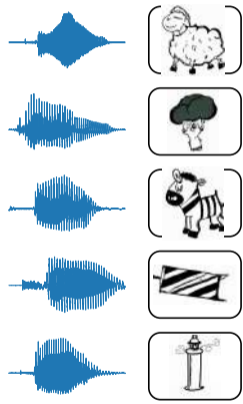
# Unimodal one-shot learning and classification



Fei-Fei et al., “One-shot learning of object categories,” *TPAMI*, 2006.  
Lake et al., “One-shot learning of generative speech concepts,” in *CogSci*, 2014.

# Multimodal one-shot learning and matching

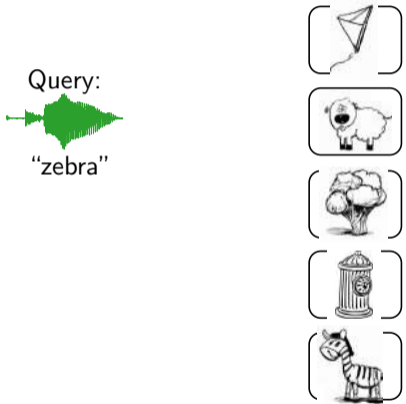
Support set



Multimodal one-shot learning

The support set consists of five rows, each containing a blue audio waveform on the left and a corresponding image on the right. The images are: a sheep, a mushroom, a zebra, a striped flag, and a lighthouse. A vertical bracket on the left side of the support set is labeled "Support set".

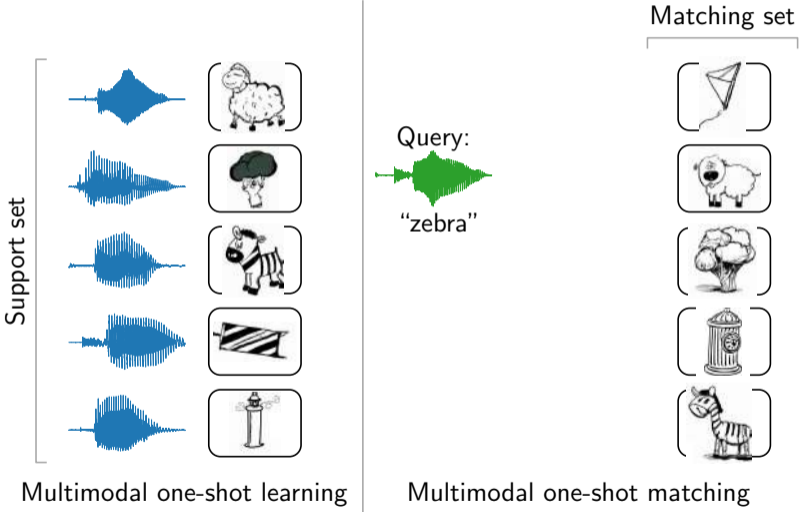
Query:



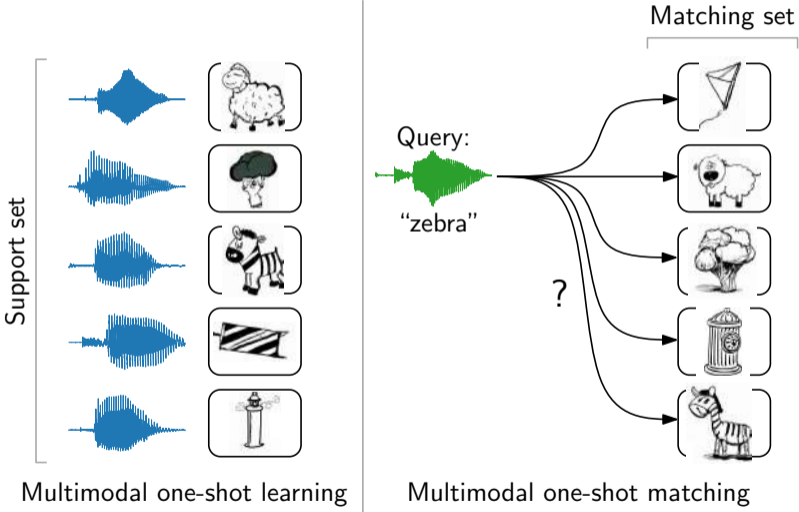
Multimodal one-shot matching

The query consists of a green audio waveform labeled "zebra" and a vertical list of five images: a kite, a sheep, a tree, a trash can, and a zebra. A vertical line separates the support set from the query.

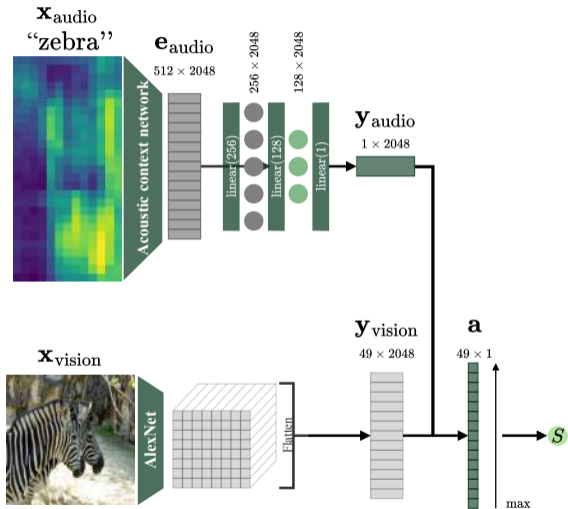
# Multimodal one-shot learning and matching



# Multimodal one-shot learning and matching



# Multimodal attention network (MattNet)



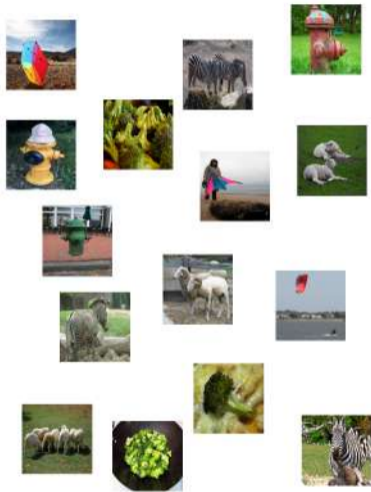
The acoustic context network is a CPC model trained on Places and LibriSpeech (level 1).

# How can we train MattNet with just a few shots?

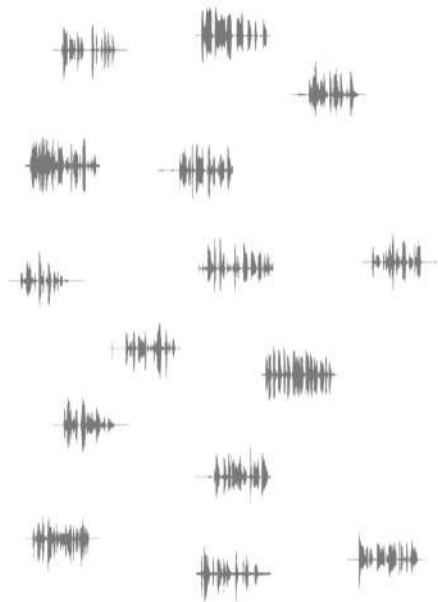


Support set

- Train on background classes
- Naively fine-tune on support-set pairs  
(Miller and Harwath, 2022)
- Use unlabelled unimodal data to artificially construct more pairs

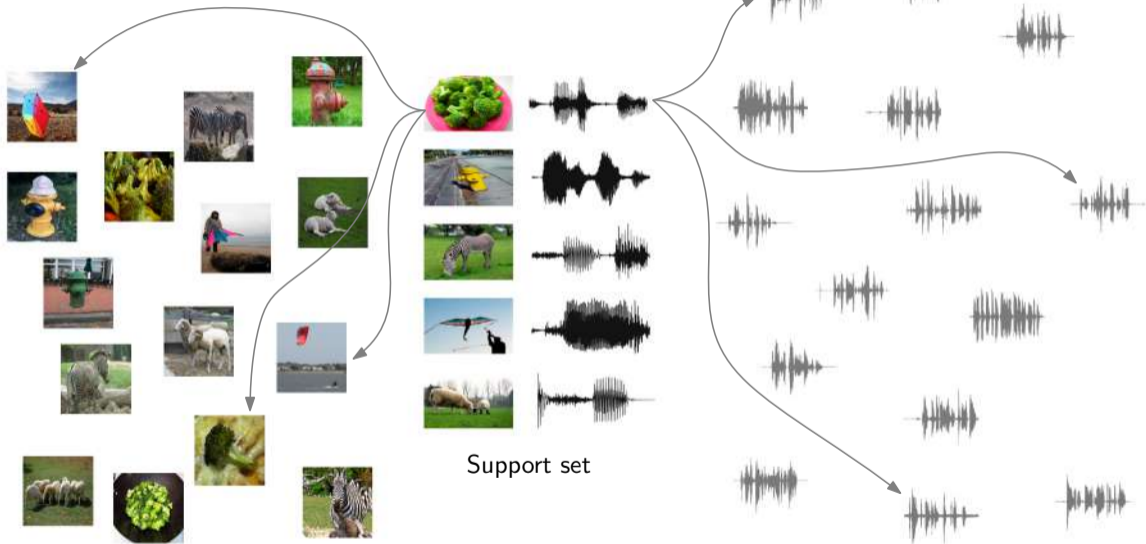


Support set

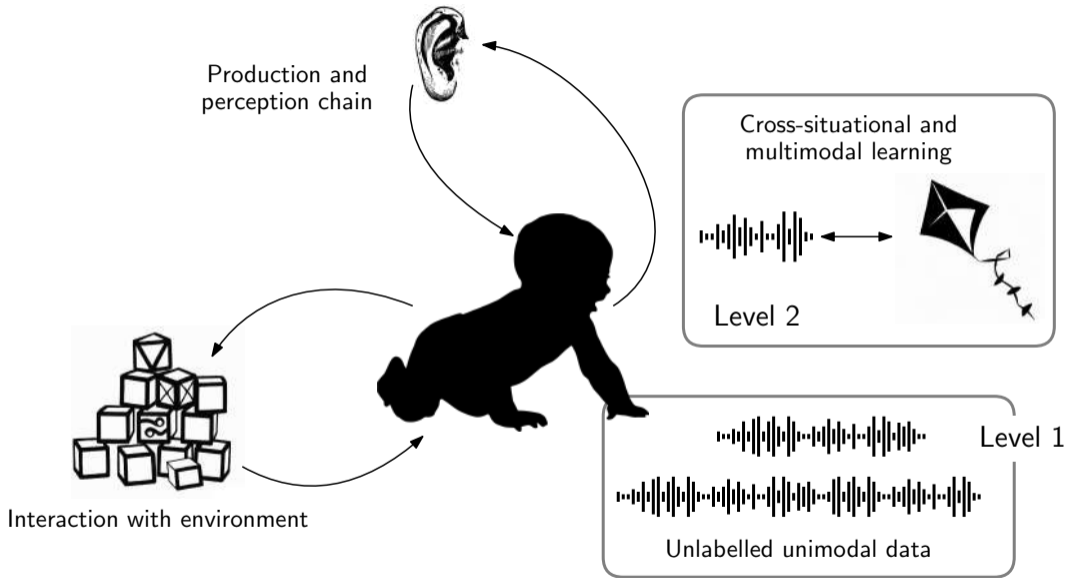












# Few-shot retrieval results

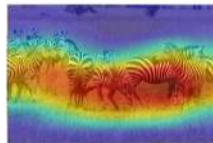
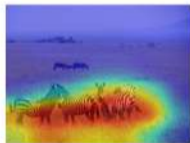
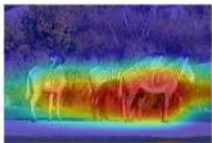
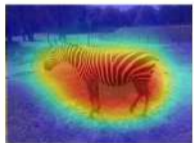
$P@N$  retrieval accuracies (%):

Model	Number of shots, $K$			
	5	10	50	100
DAVEnet (Miller and Harwath, 2022)	–	8.4±0.0	24.0±0.1	35.5±0.2
MattNet background classes	22.0±0.4	24.1±0.8	22.7±0.5	23.2±1.1
MattNet naive fine-tuned	13.2±0.6	34.8±0.7	40.9±0.3	40.5±0.5
MattNet with mining	<b>44.4±0.0</b>	<b>43.4±0.1</b>	<b>40.2±0.0</b>	42.5±0.1

# Attention visualisation



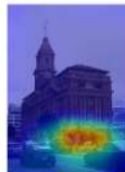
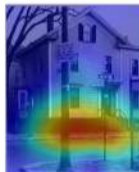
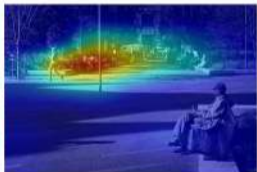
“zebra”



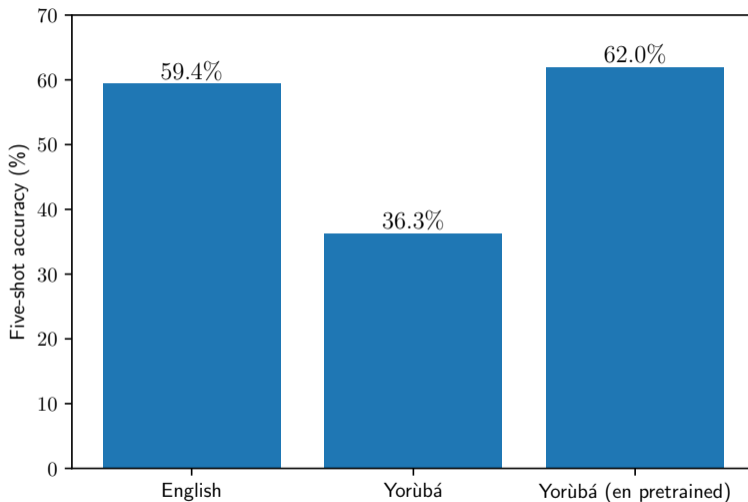
# Attention visualisation



“ fire hydrant ”

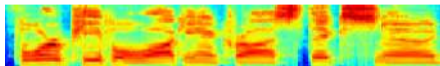
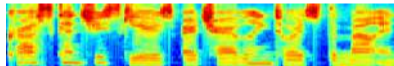
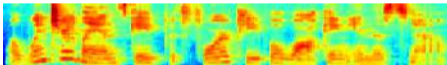


# Yorùbá few-shot classification accuracies



Nortje et al., "Visually grounded few-shot word learning in low-resource settings," *arXiv*, 2023.  
<https://www.kamperh.com/yfacc/>

# Using images for grounding speech

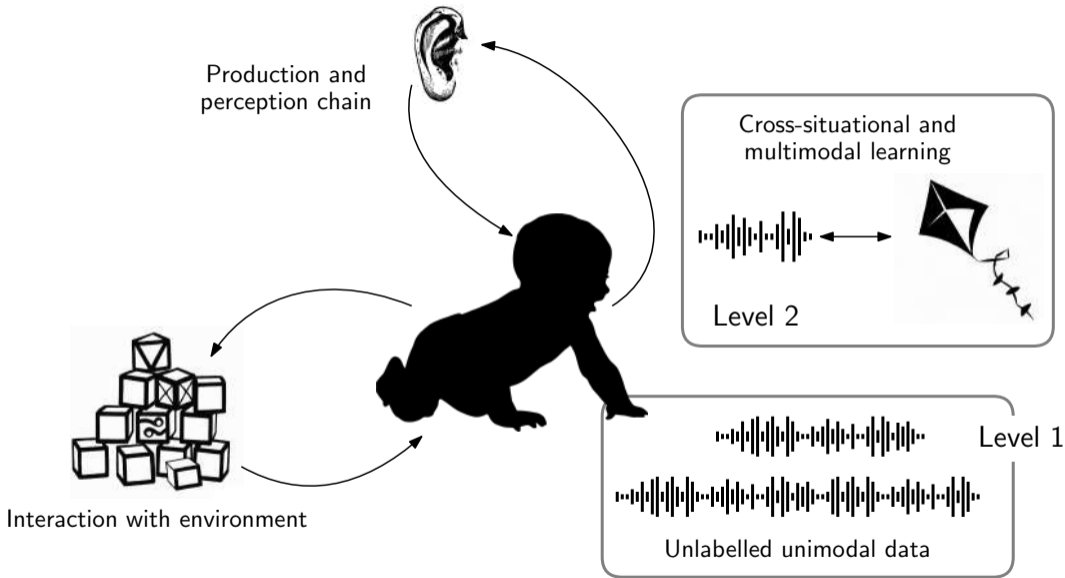




# Many remaining questions

- Catastrophic forgetting (Miller and Harwath, 2022)
- Cognitive plausibility and what this actually tells us about cognition
- Shortcomings in the mining approach
- Explore this to investigate the mutual exclusivity bias

Would love to get your inputs!



## 2. Probing self-supervised speech models by listening



Benjamin  
van Niekerk



Matthew  
Baas



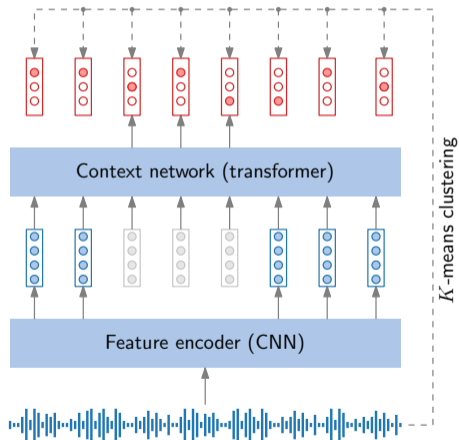
Marc-André  
Carbonneau

Baas et al., "Voice conversion with just nearest neighbors," in *Interspeech*, 2023.

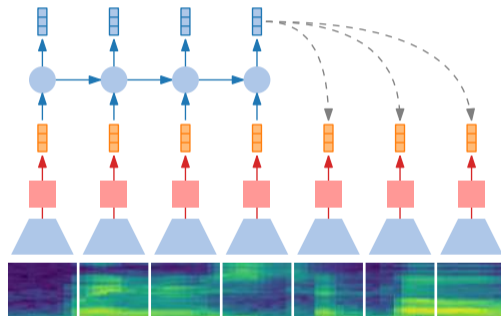
van Niekerk et al., "Rhythm modeling for voice conversion," *IEEE SPL*, 2023.

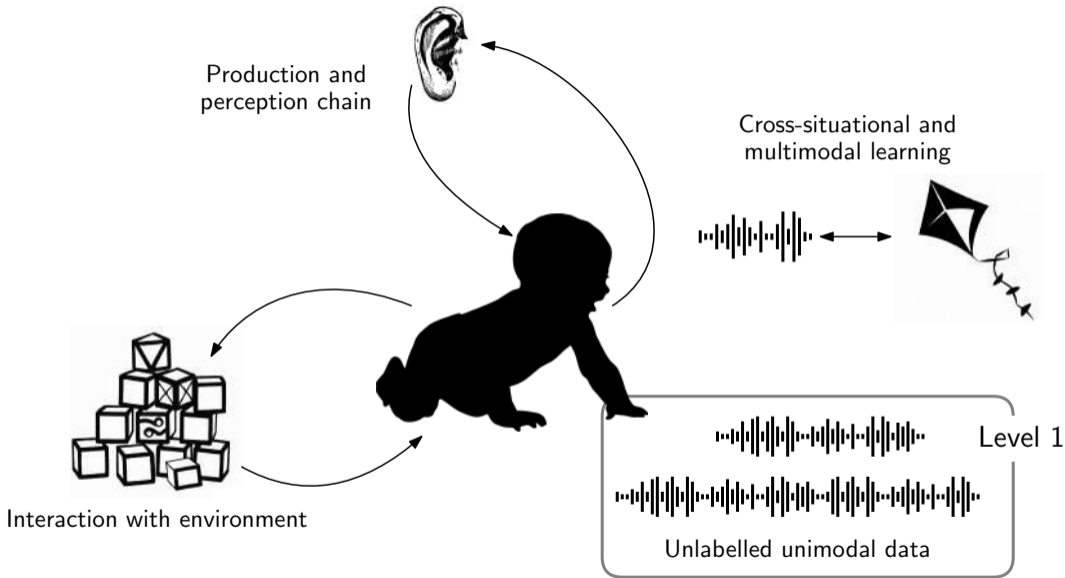
# Self-supervised speech models

HuBERT / WavLM:



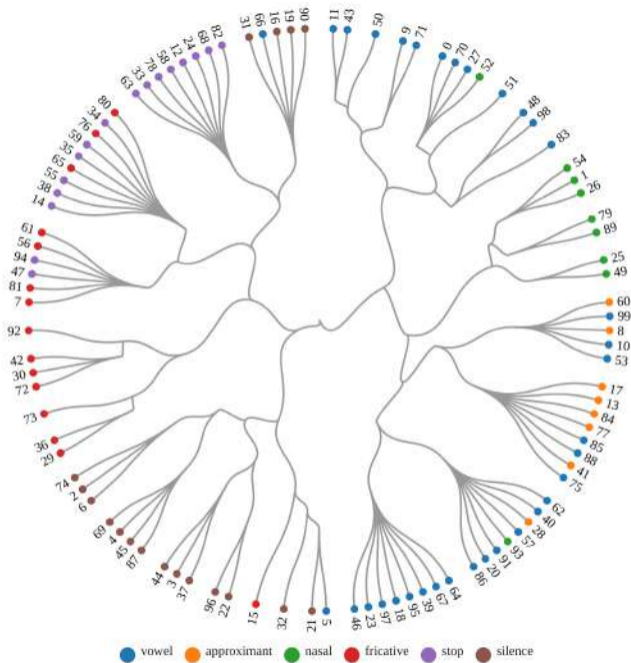
Contrastive predictive coding (CPC):

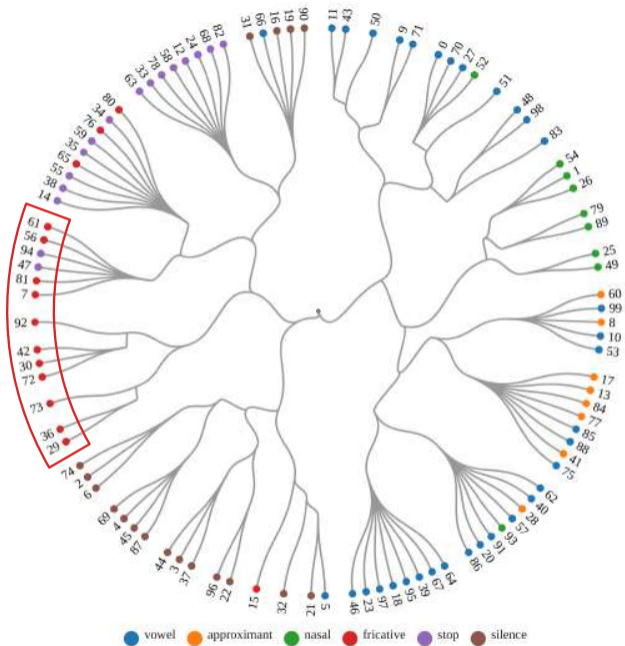




We use voice alteration and voice conversion as a probe to show you how phonetic content and speaker are captured.

(But it's really just an excuse . . .)





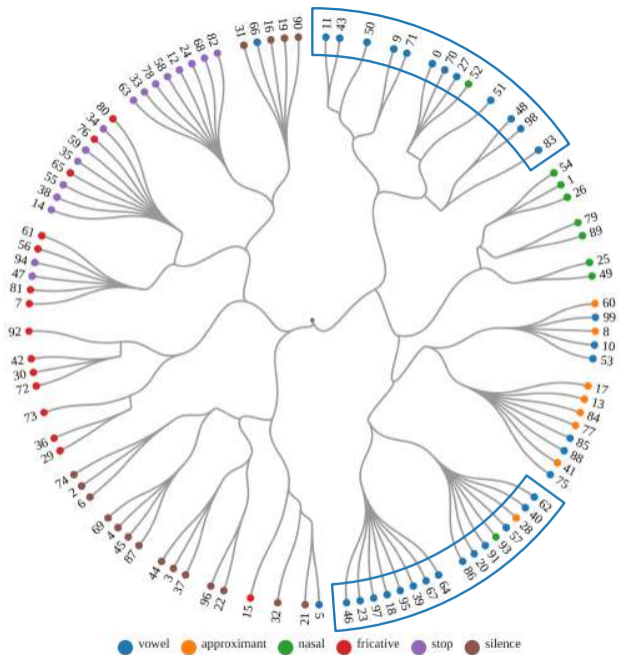
No modification:

Play

Fricatives:

Play



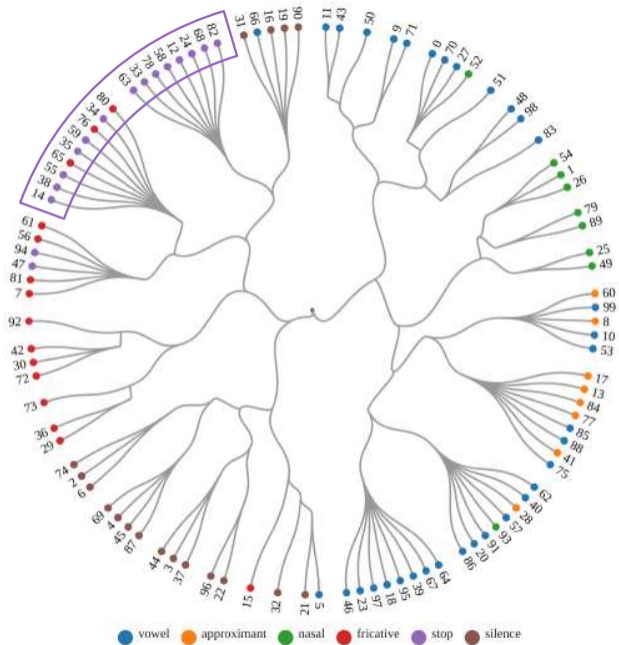


No modification:

Play

Vowels:

Play

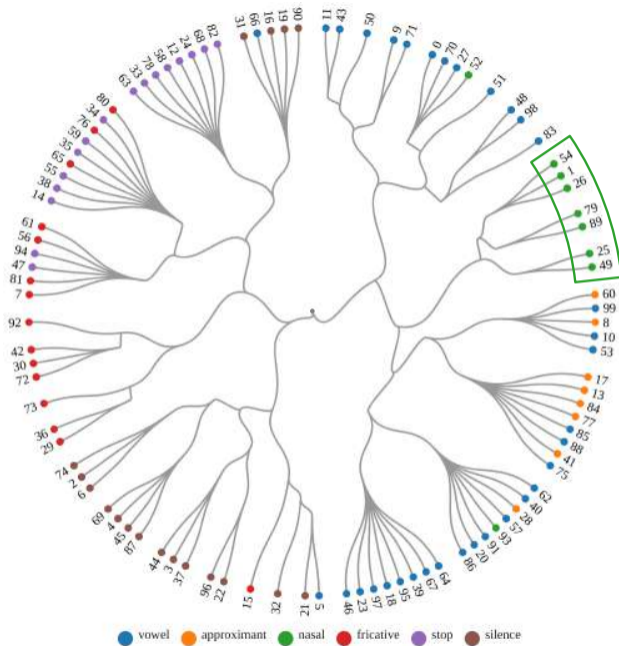


No modification:

Play

Stops:

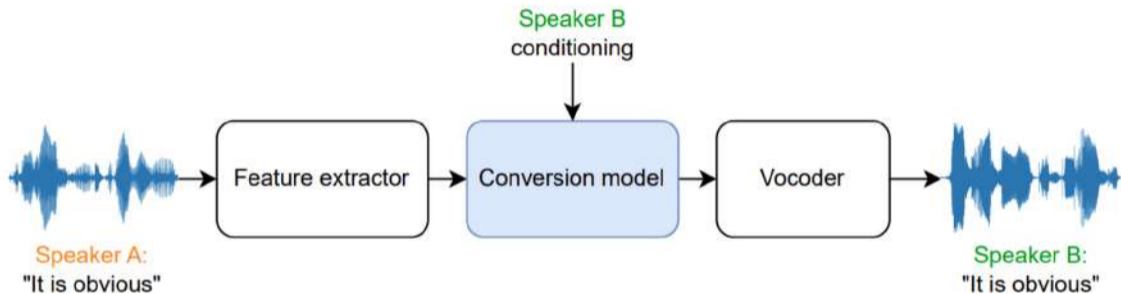
Play



No modification: [Play](#)

Nasals: [Play](#)

# Voice conversion

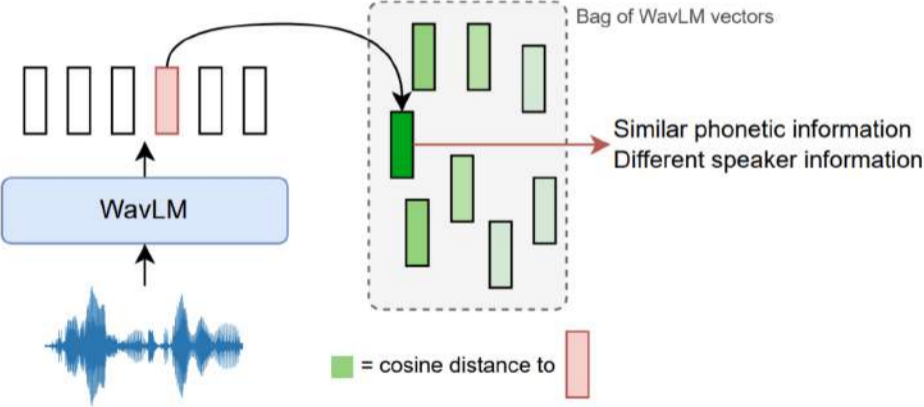


Source: [Play](#)

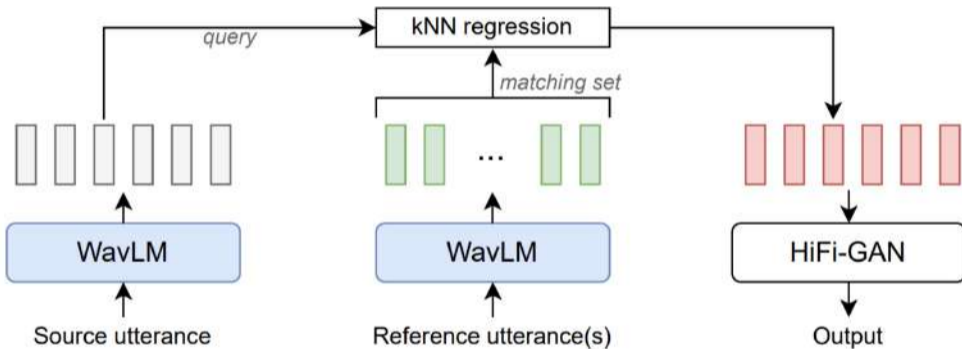
Reference: [Play](#)

Output: [Play](#)

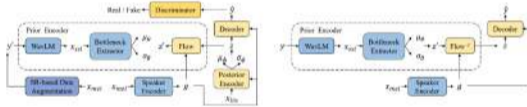
# Our key idea



# $k$ -nearest neighbours voice conversion (kNN-VC)



# Existing voice conversion systems



FreeVC [2022]

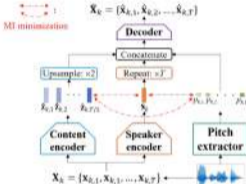
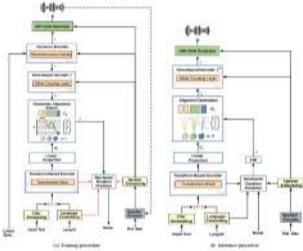


Figure 1: Diagram of the proposed VQMIVC system.

VQMIVC [2021]



YourTTS [2023]

# Voice conversion results

Model	WER ↓	EER ↑	MOS ↑	SIM ↑
<i>Testset topline</i>	5.96	–	4.24	3.19
VQMIVC (Wang et al., 2021)	59.46	2.22	2.70	2.09
YourTTS (Casanova et al., 2022)	11.93	25.32	3.53	2.57
FreeVC (Li et al., 2022)	7.61	8.97	<b>4.07</b>	2.38
kNN-VC	<b>7.36</b>	<b>37.15</b>	<b>4.03</b>	<b>2.91</b>




# Fun samples


Cross-lingual conversion:

Source: 


Reference: 

Output: 


Whispered music conversion:

Source: 


Reference: 

Output: 

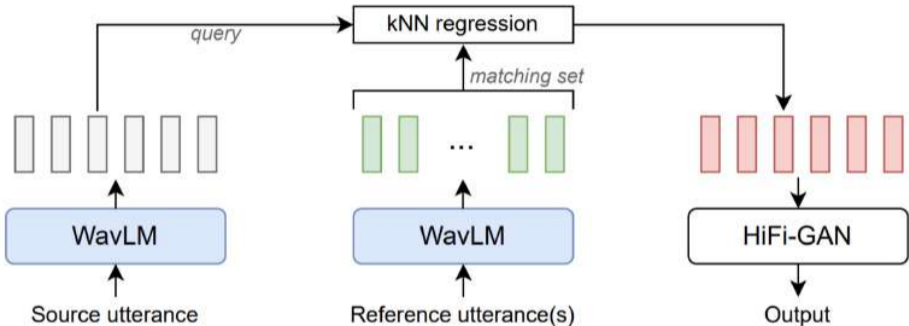
Human-to-animal conversion:

Source: 

Reference: 

Output: 

# Voice conversion with stuttered reference speech



Source: [Play](#)

Reference: [Play](#)

Output: [Play](#)

Baseline: [Play](#) (TTS)

Source: [Play](#)

Reference: [Play](#)

Output: [Play](#)

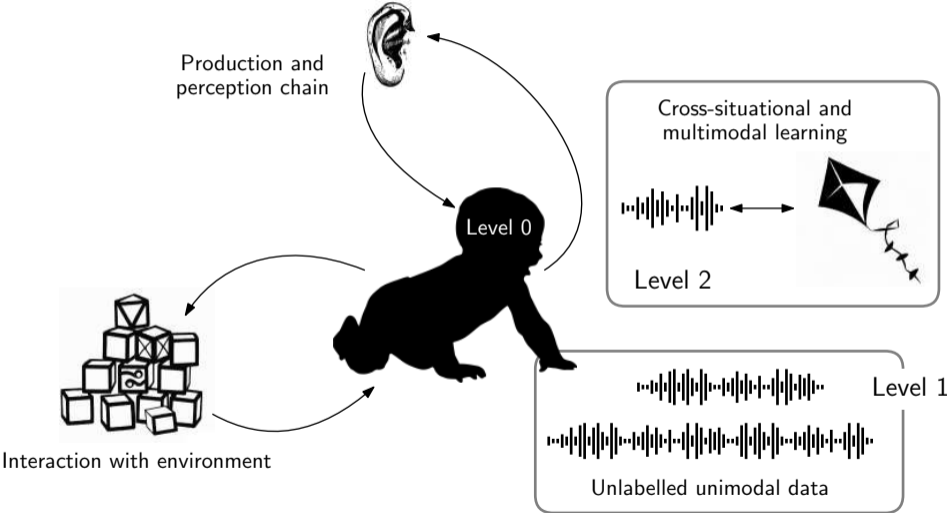
Baseline: [Play](#) (manual)

# What does this tell us about self-supervised speech models?

- Broader phonetic categories are captured in hierarchy
- Phonetic content is matched through cosine distance
- But speaker characteristics are also still strongly captured

All of this is kind of expected, but it is still cool to be able to hear it!

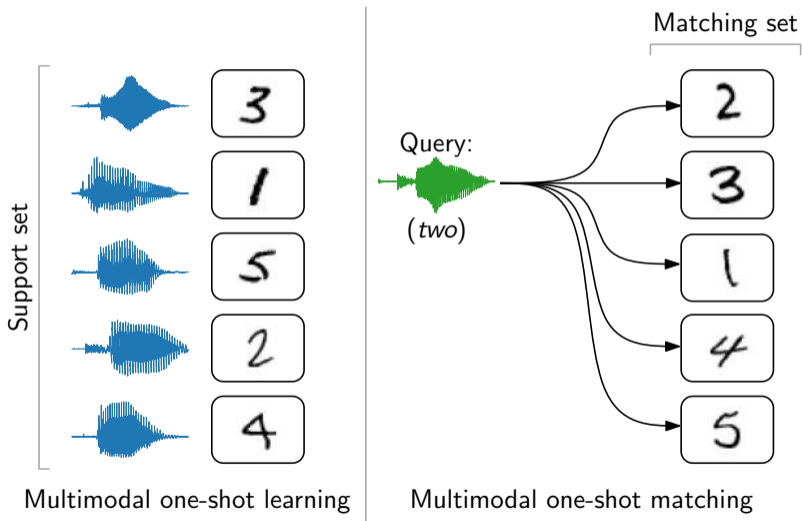
# Conclusion



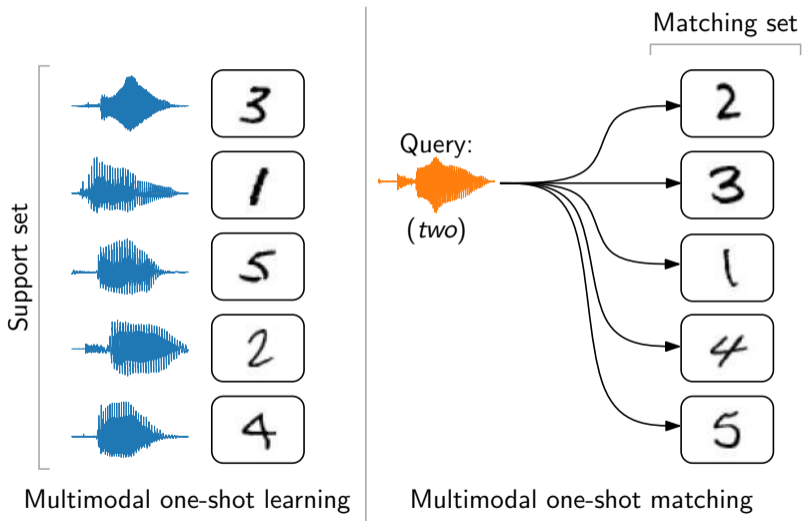
<https://bshall.github.io/knn-vc/>

<https://www.kamperh.com/>

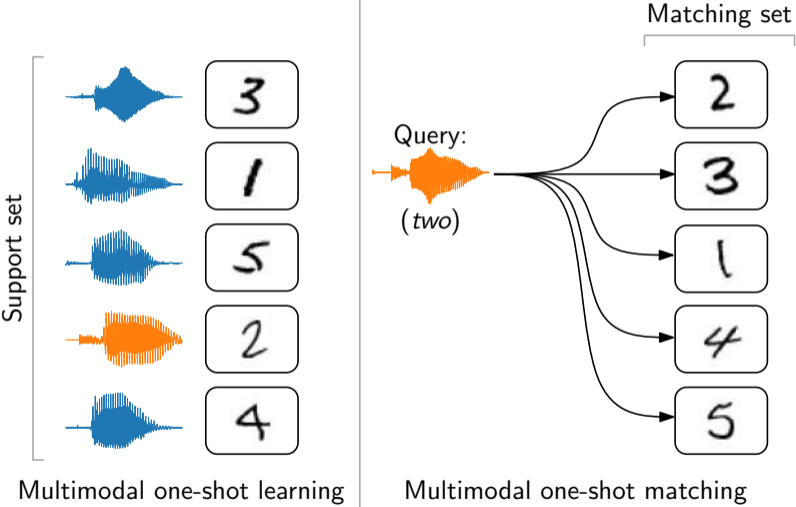
# Two-step (indirect) multimodal one-shot approach



# Two-step (indirect) multimodal one-shot approach

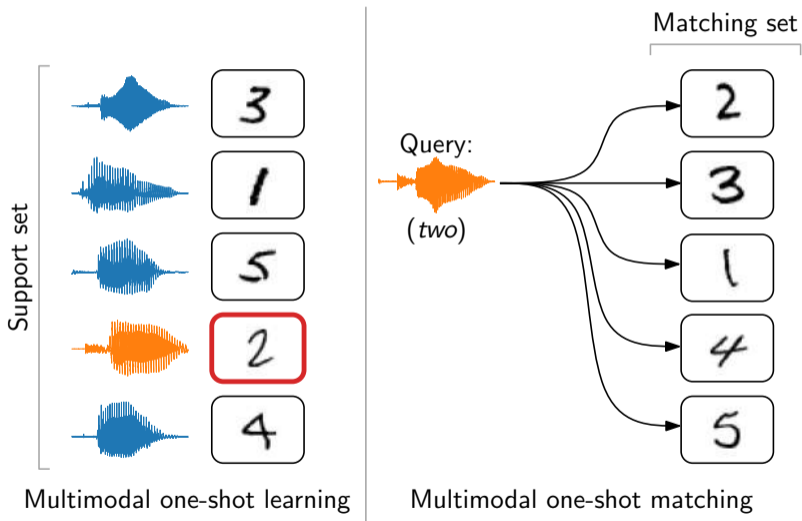


# Two-step (indirect) multimodal one-shot approach

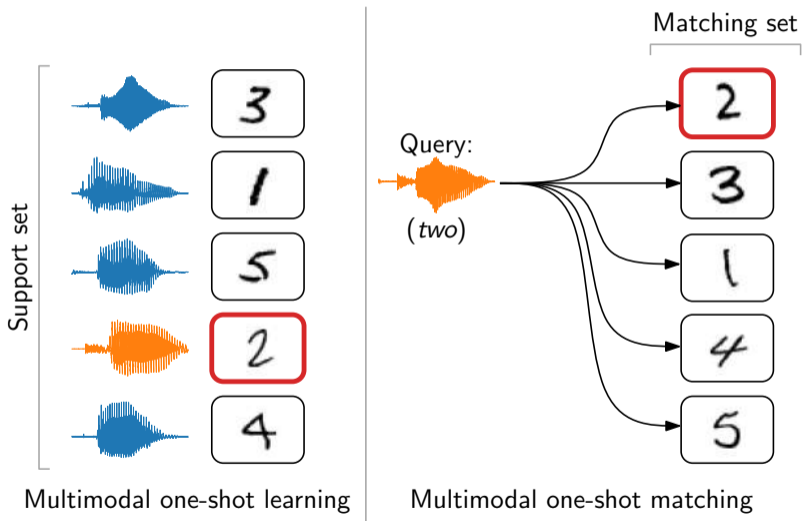




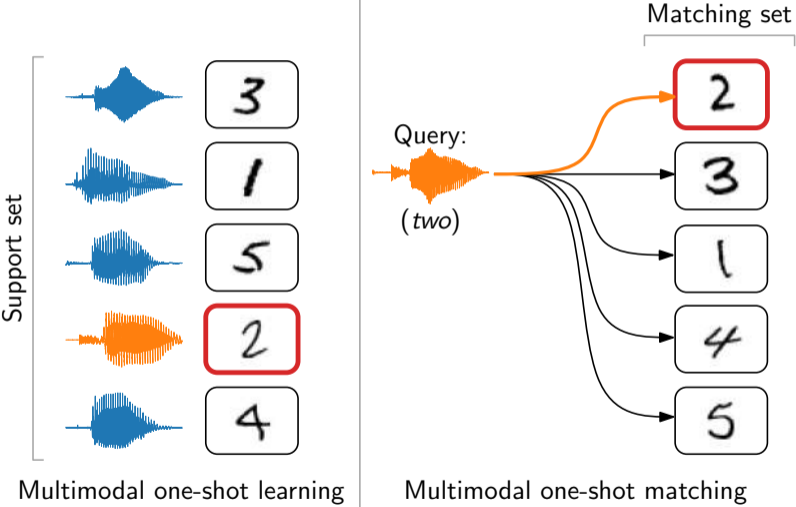
# Two-step (indirect) multimodal one-shot approach



# Two-step (indirect) multimodal one-shot approach



# Two-step (indirect) multimodal one-shot approach



# Two-step (indirect) multimodal one-shot approach

