Speech systems that emulate human language acquisition

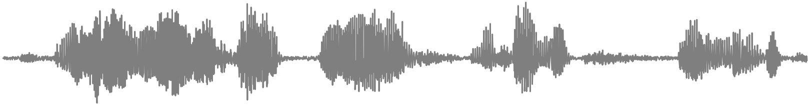
SLS group, MIT, Apr. 2024

Herman Kamper

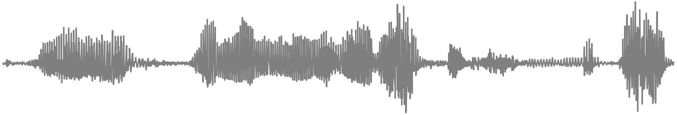
E&E Engineering, Stellenbosch University, South Africa

http://www.kamperh.com/

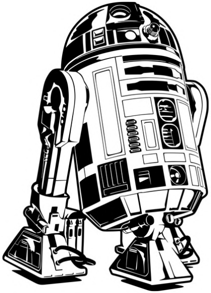
Supervised speech recognition and synthesis

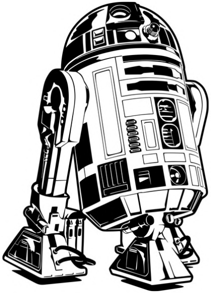


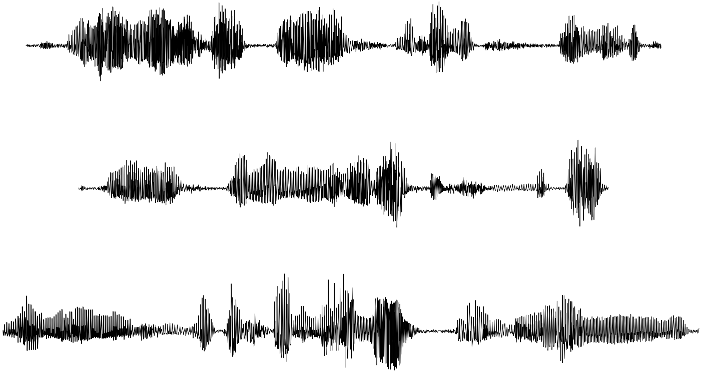
i had to think of some example speech

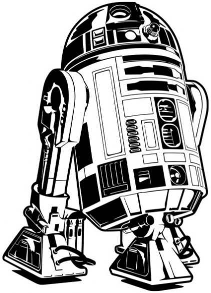


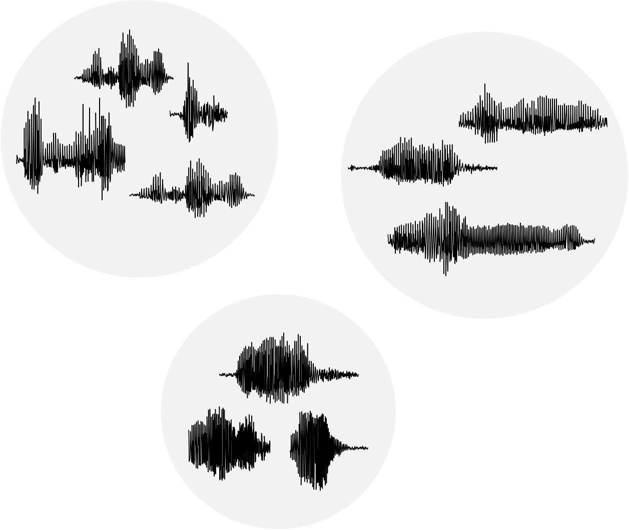
since speech recognition is really cool

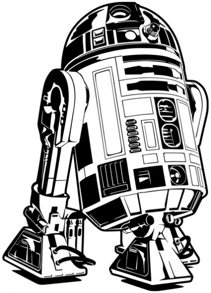


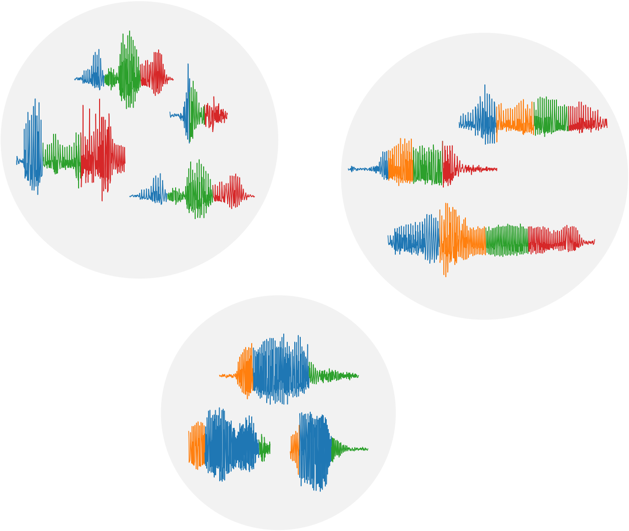










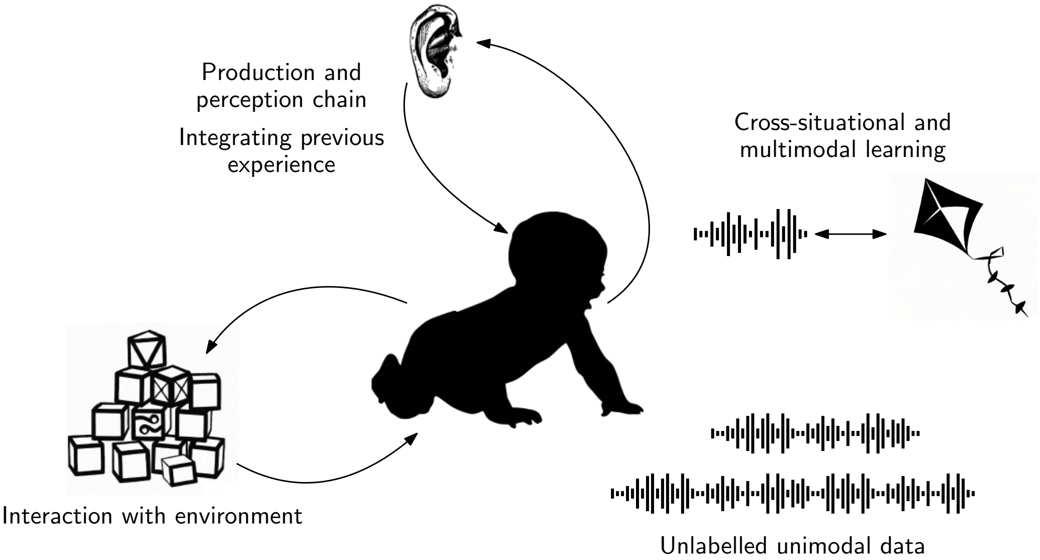


Why attempt to emulate language acquisition?

Improvements in speech technology

New insights and approaches for machines that learn

New insights into human learning



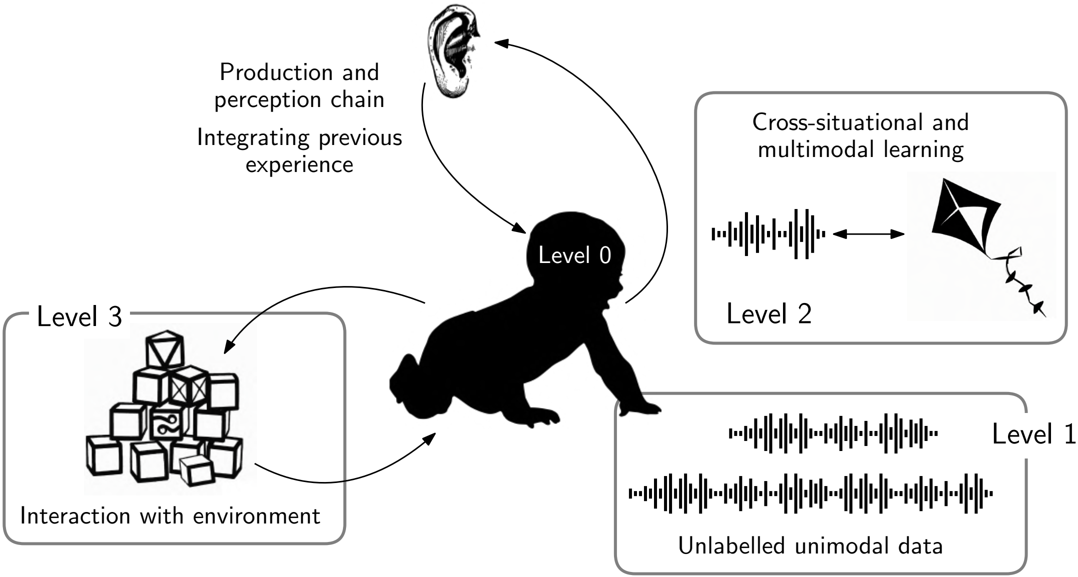
Production and perception chain

Integrating previous experience

Cross-situational and multimodal learning

Interaction with environment

Unlabelled unimodal data



Production and perception chain

Integrating previous experience

Level 0

Cross-situational and multimodal learning

Level 2

Level 3

Interaction with environment

Level 1

Unlabelled unimodal data

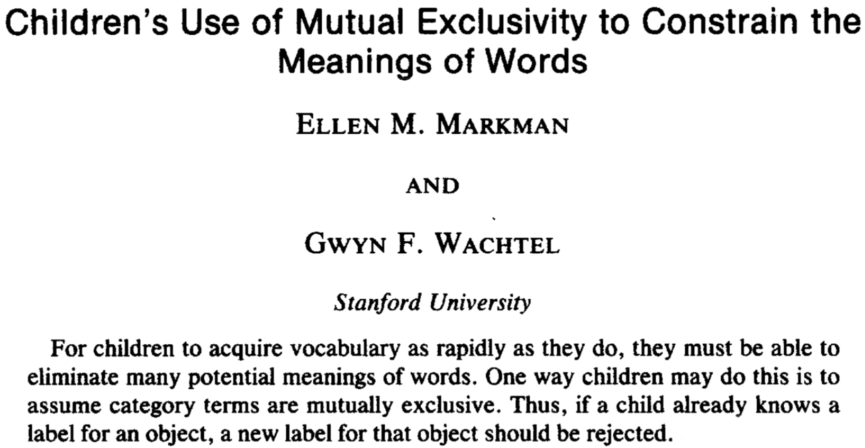
1. Mutual exclusivity in visually grounded speech models

Leanne
Nortje

Dan
Oneață

Yevgen
Matusevych

Nortje et al., "Visually grounded few-shot word acquisition with fewer shots," in *Interspeech*, 2023.
Nortje et al., "Visually grounded speech models have a mutual exclusivity bias," *Accepted*, 2024.



Children's Use of Mutual Exclusivity to Constrain the Meanings of Words

ELLEN M. MARKMAN

AND

GWYN F. WACHTEL

Stanford University

For children to acquire vocabulary as rapidly as they do, they must be able to eliminate many potential meanings of words. One way children may do this is to assume category terms are mutually exclusive. Thus, if a child already knows a label for an object, a new label for that object should be rejected.

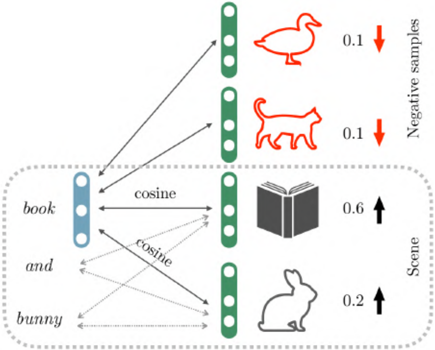
A

B

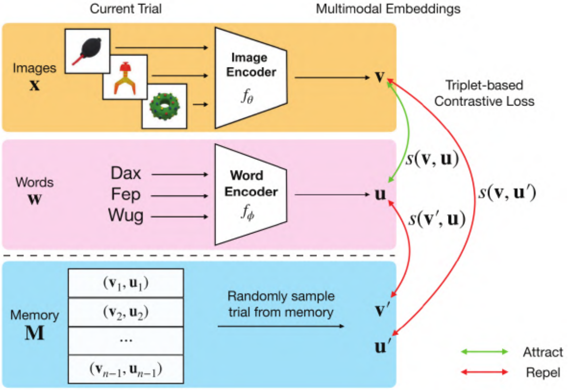
?

Previous computational studies

Previous computational studies



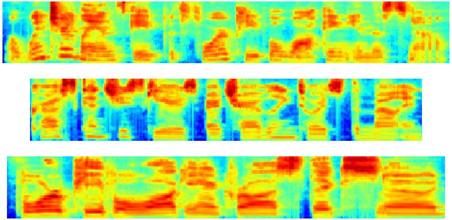
(Gulordava et al., 2020)



Current Trial
Multimodal Embeddings
Images
$\mathbf{x}$
Image Encoder
f_θ
$\mathbf{v}$
Triplet-based Contrastive Loss
Words
$\mathbf{w}$
Dax
Fep
Wug
Word Encoder
f_ϕ
$\mathbf{u}$
$s(\mathbf{v},\mathbf{u})$
$s(\mathbf{v},\mathbf{u}')$
$s(\mathbf{v}',\mathbf{u})$
Memory
$\mathbf{M}$
$(\mathbf{v}_1,\mathbf{u}_1)$
$(\mathbf{v}_2,\mathbf{u}_2)$
...
$(\mathbf{v}_{n-1},\mathbf{u}_{n-1})$
Randomly sample trial from memory
$\mathbf{v}'$
$\mathbf{u}'$
Attract
Repel

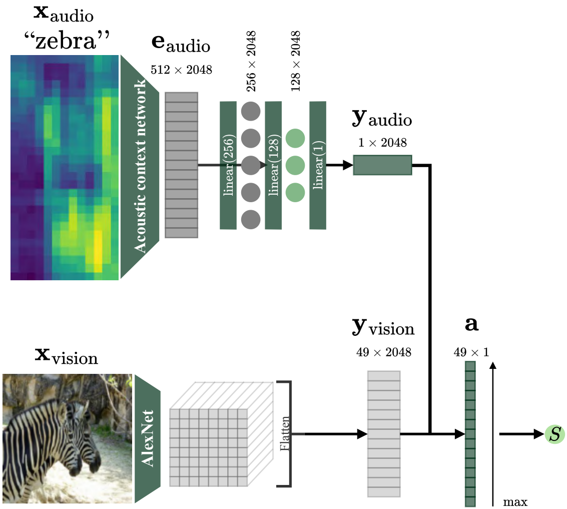
(Vong and Lake, 2022)

Visually grounded speech models



Harwath et al., "Unsupervised learning of spoken language with visual context," in *NeurIPS*, 2016.

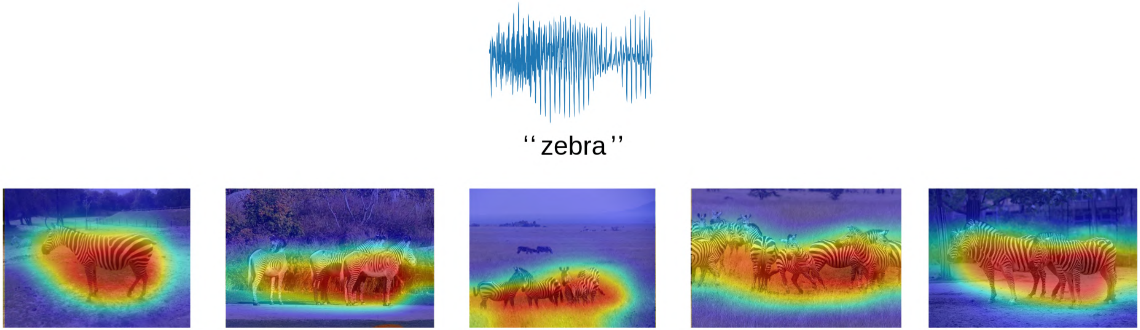
Multimodal attention network (MattNet)



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

The vision branch is intialised with a self-supervised variant of AlexNet (level 1).

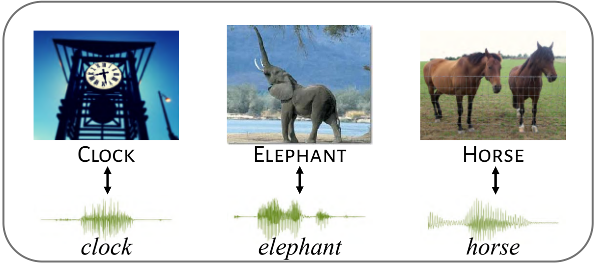
Attention visualisation



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

Testing a visually grounded speech model for the ME bias

Given during training:

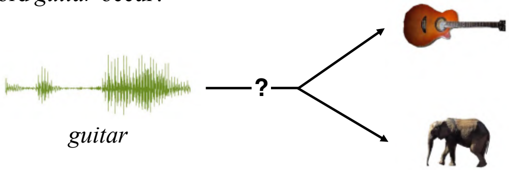


CLOCK ↕ *clock*

ELEPHANT ↕ *elephant*

HORSE ↕ *horse*

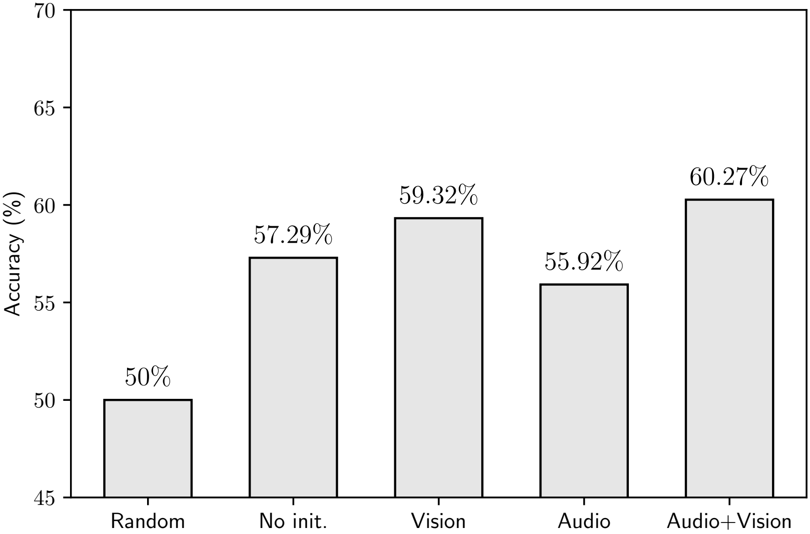
Test-time question: In which picture does the novel spoken keyword *guitar* occur?

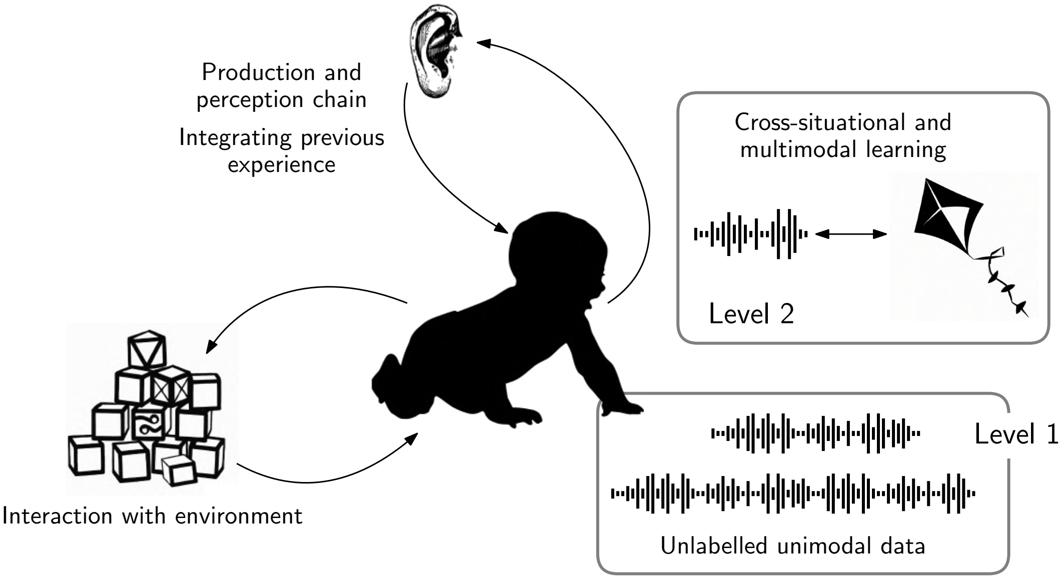


guitar

?

Mutual exclusivity results





Production and perception chain

Integrating previous experience

Cross-situational and multimodal learning

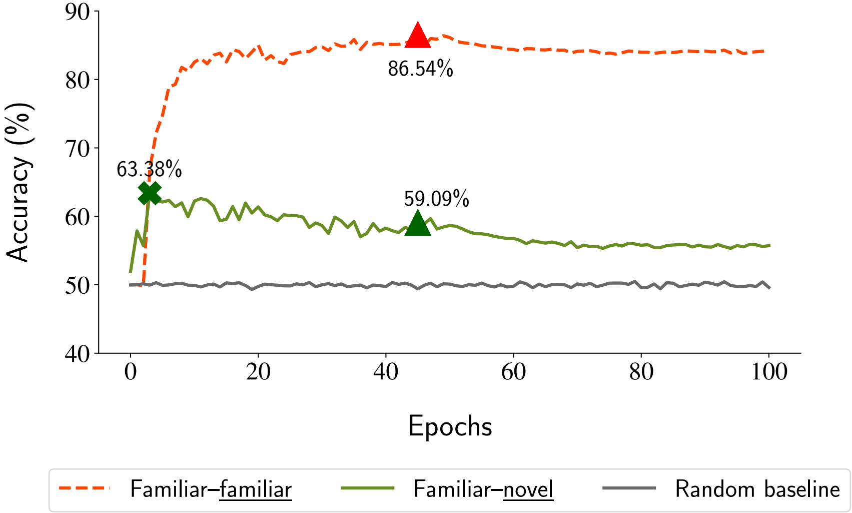
Level 2

Level 1

Unlabelled unimodal data

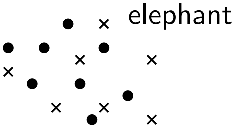
Interaction with environment

Mutual exclusivity results



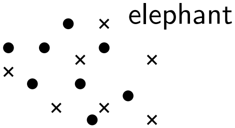
How is the representation space organised?

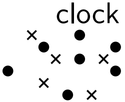
How is the representation space organised?



*This is just a cartoon!

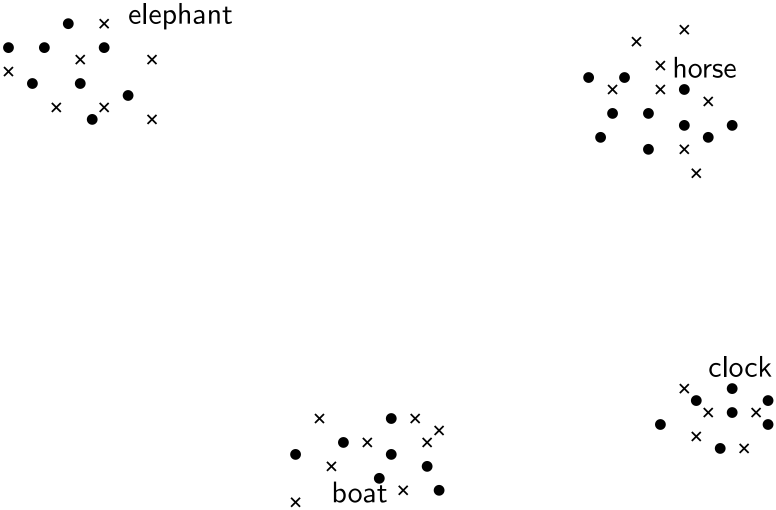
How is the representation space organised?





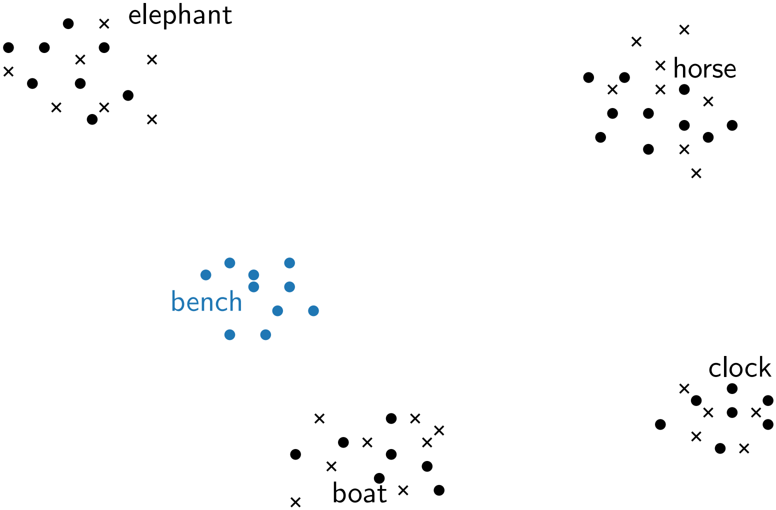
*This is just a cartoon!

How is the representation space organised?



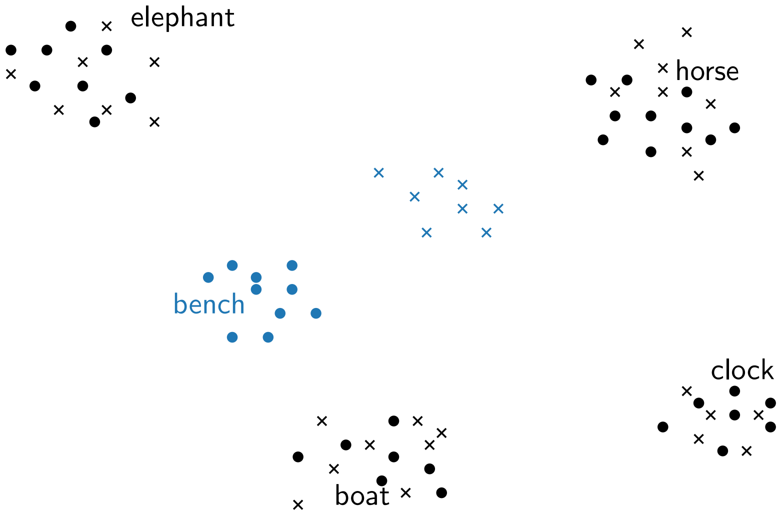
*This is just a cartoon!

How is the representation space organised?



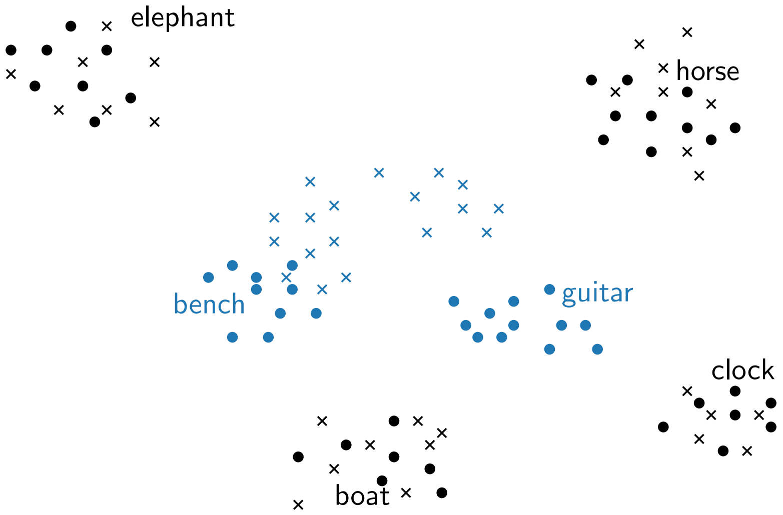
*This is just a cartoon!

How is the representation space organised?



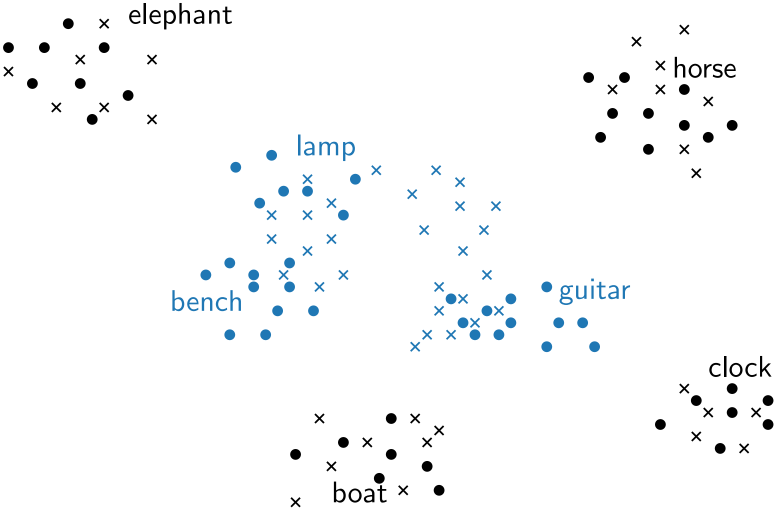
*This is just a cartoon!

How is the representation space organised?



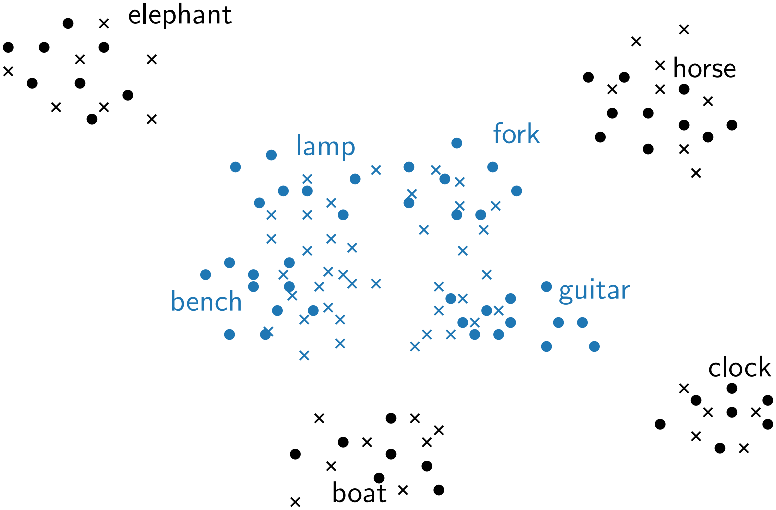
*This is just a cartoon!

How is the representation space organised?



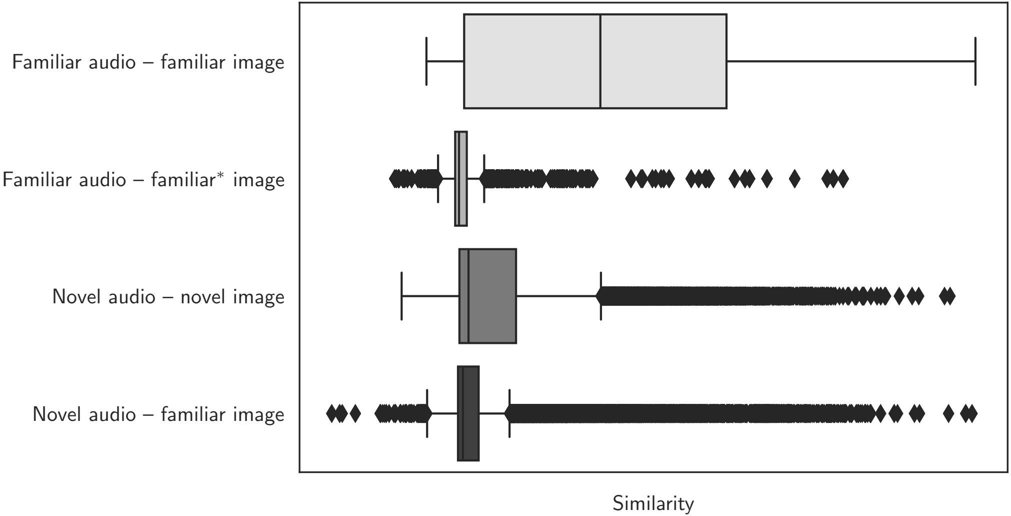
*This is just a cartoon!

How is the representation space organised?



*This is just a cartoon!

How is the representation space organised?

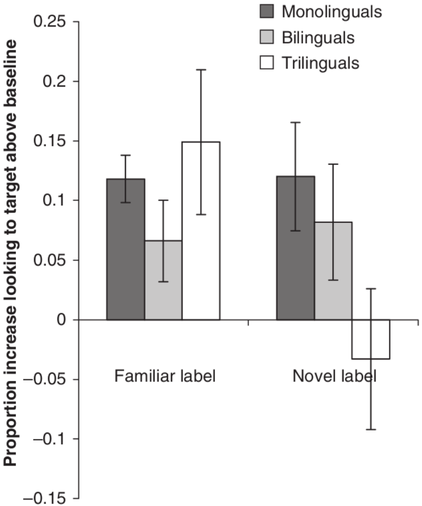


Conclusions and future work

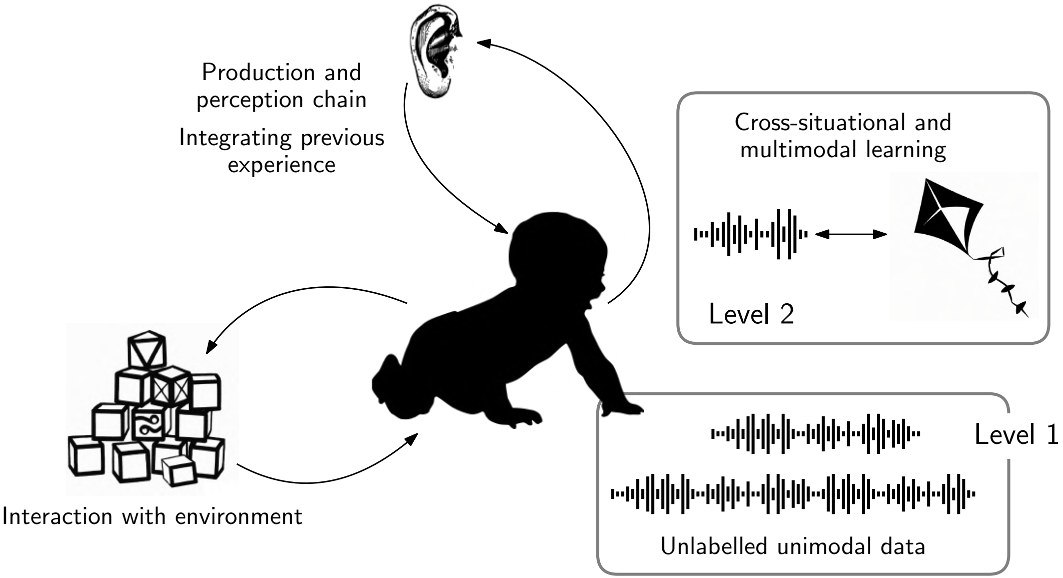
- Showed an example of how we can compare an artificial learner to human infants
- Use speech and not written words
- Adds weight that visually grounded speech model could be studied as a cognitive proxy

Conclusions and future work

- Showed an example of how we can compare an artificial learner to human infants
- Use speech and not written words
- Adds weight that visually grounded speech model could be studied as a cognitive proxy
- Future work: Mutual exclusivity in multilingual models



(Byers-Heinlein and Werker, 2009)



Production and perception chain

Integrating previous experience

Interaction with environment

Cross-situational and multimodal learning

Level 2

Level 1

Unlabelled unimodal data

2. Probing self-supervised speech models by listening

Benjamin
van Niekerk

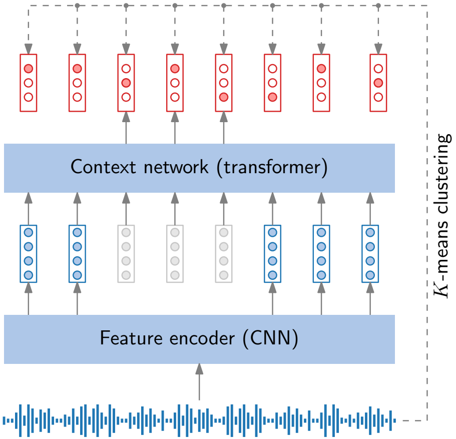
Matthew
Baas

Marc-André
Carbonneau

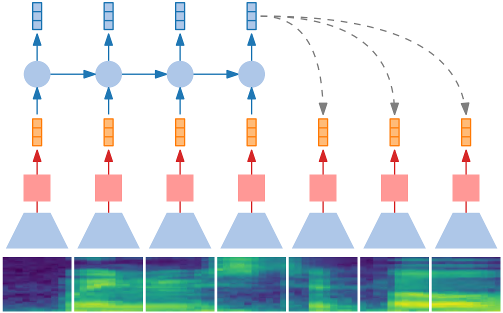
Baas et al., "Voice conversion with just nearest neighbors," in *Interspeech*, 2023.
van Niekerk et al., "Rhythm modeling for voice conversion," *IEEE SPL*, 2023.

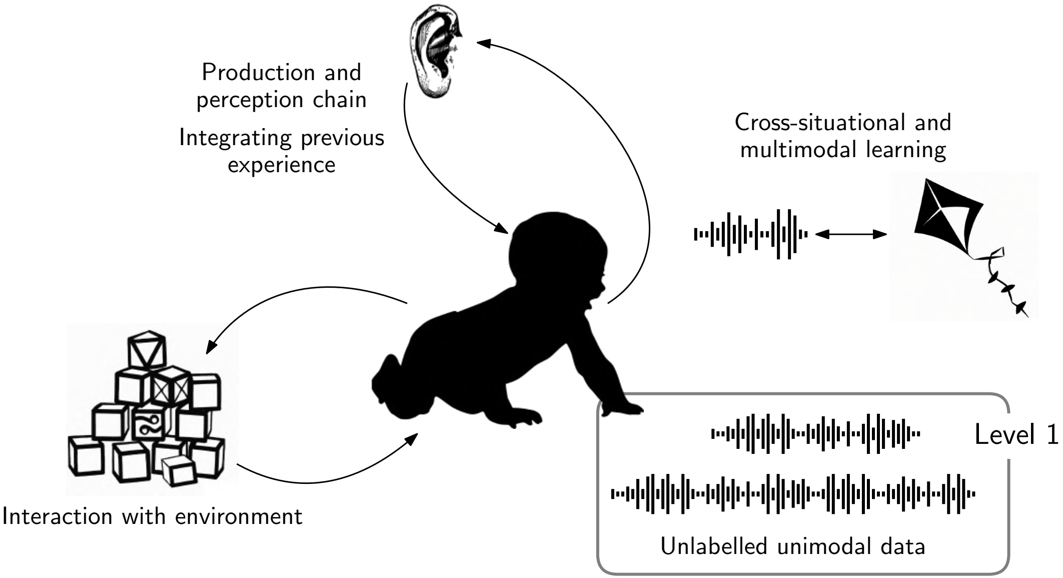
Self-supervised spoken language models

HuBERT / WavLM:



Contrastive predictive coding (CPC):





Production and perception chain

Integrating previous experience

Cross-situational and multimodal learning

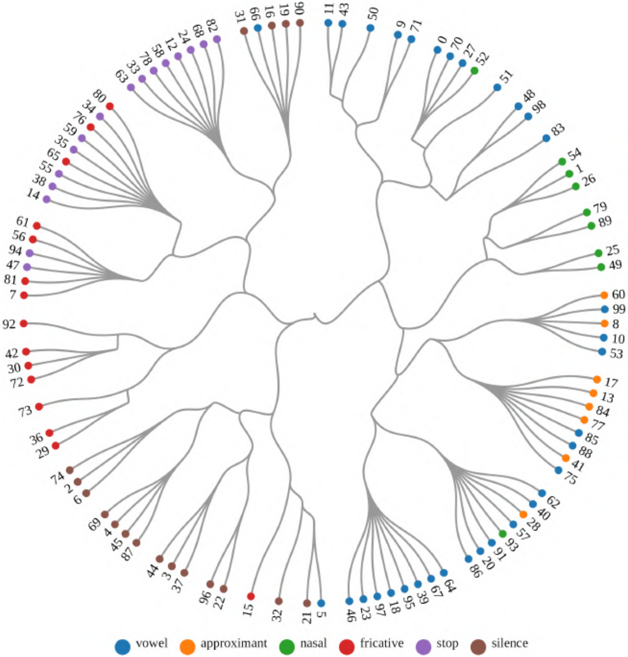
Interaction with environment

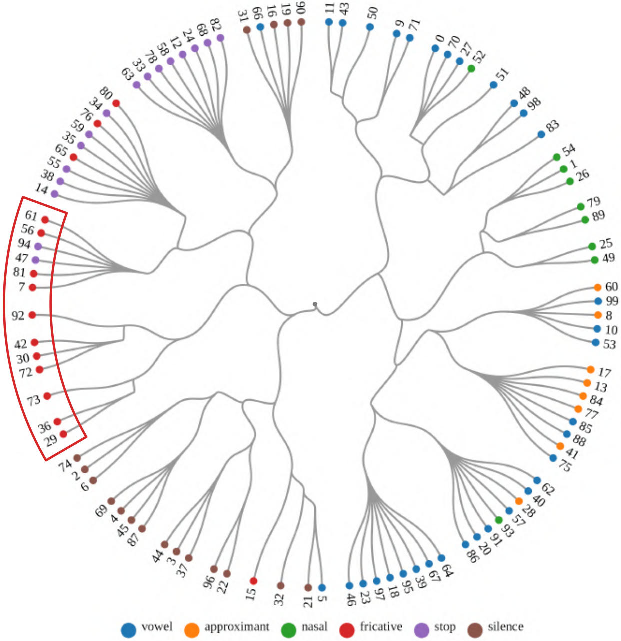
Level 1

Unlabelled unimodal data

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

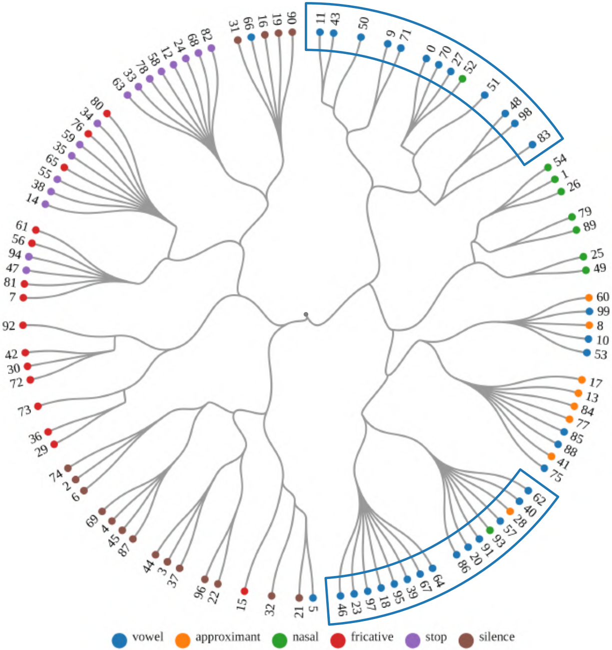
(But it's really just an excuse . . .)


31
66
16
19
90
11
43
50
9
71
0
70
27
52
51
48
98
83
54
1
26
79
89
25
49
60
99
8
10
53
17
13
84
77
85
88
41
75
62
40
28
57
93
91
20
86
64
67
39
95
18
97
23
46
5
21
32
15
22
96
37
3
44
87
45
4
69
6
2
74
29
36
73
72
30
42
92
7
81
47
94
56
61
14
38
55
65
35
59
76
34
80
63
33
78
58
12
24
68
82
● vowel
● approximant
● nasal
● fricative
● stop
● silence



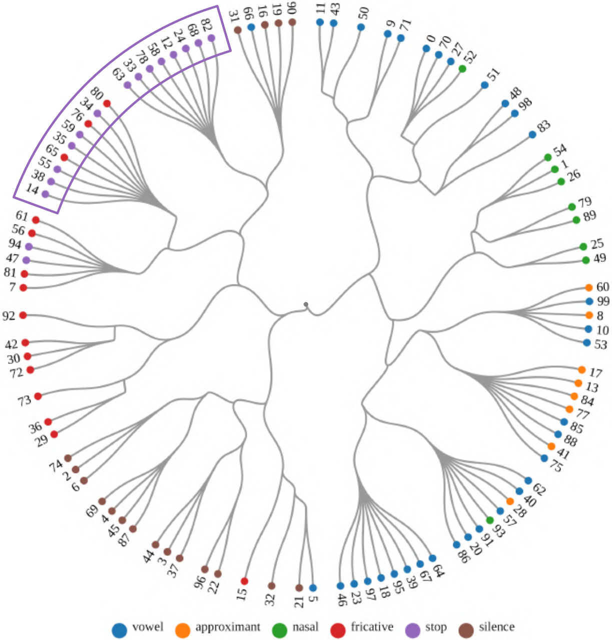
No modification: Play

Fricatives: Play



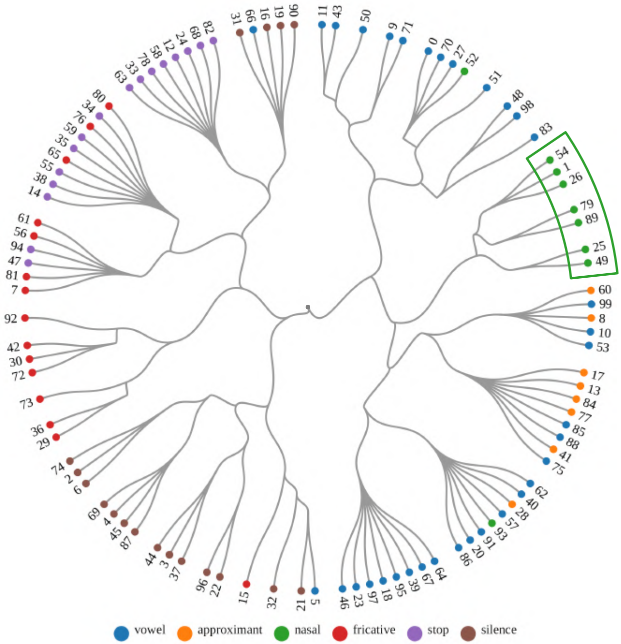
No modification: Play

Vowels: Play



No modification: Play

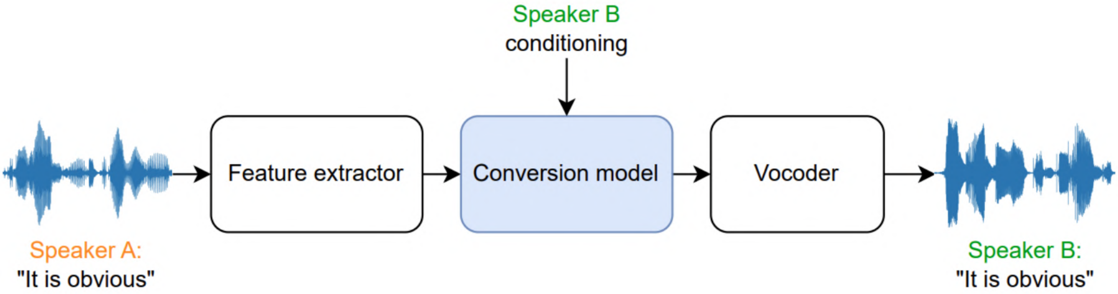
Stops: Play



No modification: Play

Nasals: Play

Voice conversion

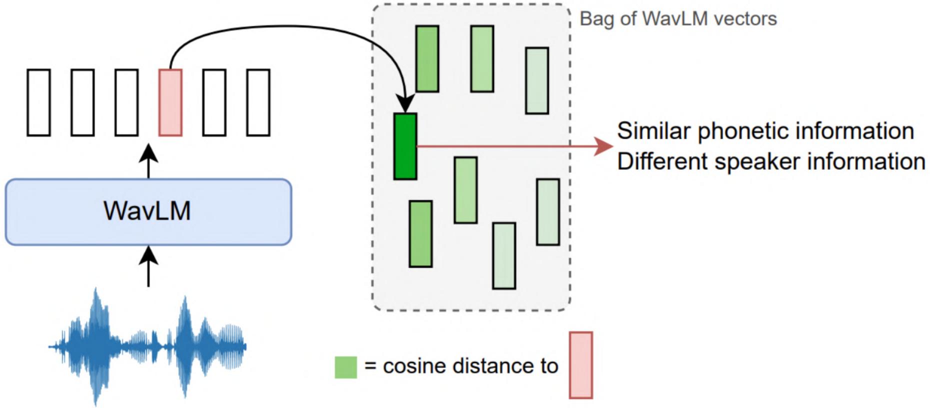


Source: Play

Reference: Play

Output: Play

Our key idea



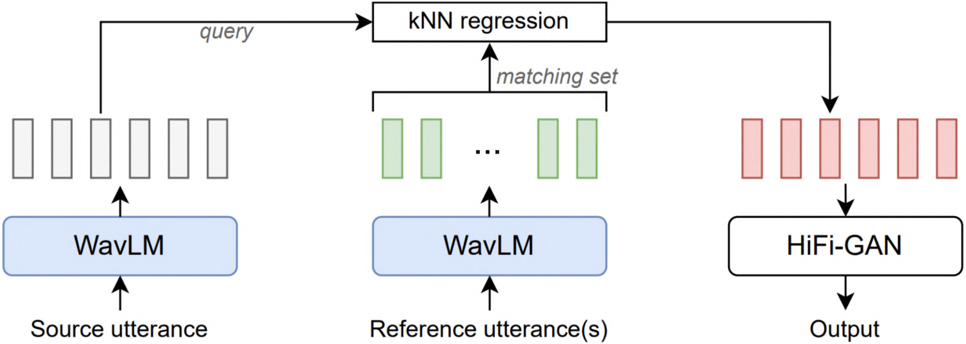
Bag of WavLM vectors

WavLM

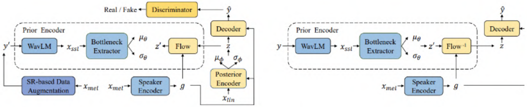
Similar phonetic information
Different speaker information

🟩 = cosine distance to 🟥

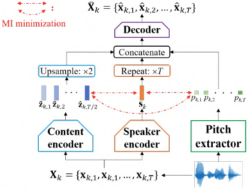
k-nearest neighbours voice conversion (kNN-VC)



Existing voice conversion systems



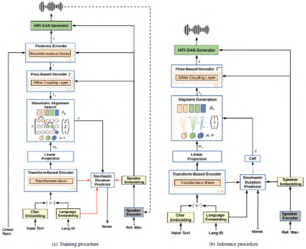
FreeVC [2022]



MI minimization
$\hat{\mathbf{X}}_k = \{\hat{\mathbf{x}}_{k,1}, \hat{\mathbf{x}}_{k,2}, \dots, \hat{\mathbf{x}}_{k,T}\}$
Decoder
Concatenate
Upsample: ×2
Repeat: ×T
$\hat{\mathbf{z}}_{k,1}$ $\hat{\mathbf{z}}_{k,2}$ $\hat{\mathbf{z}}_{k,T/2}$
$\mathbf{s}_k$
$p_{k,1}$ $p_{k,2}$ $p_{k,T}$
Content encoder
Speaker encoder
Pitch extractor
$\mathbf{X}_k = \{\mathbf{x}_{k,1}, \mathbf{x}_{k,1}, \dots, \mathbf{x}_{k,T}\}$

Figure 1: *Diagram of the proposed VQMIVC system.*

VQMIVC [2021]



YourTTS [2023]

Voice conversion results

Model	WER ↓	EER ↑	MOS ↑	SIM ↑
Testset topline	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	**4.07**	2.38
kNN-VC	**7.36**	**37.15**	**4.03**	**2.91**

Fun samples

Cross-lingual conversion:

Source: Play Reference: Play Output: Play

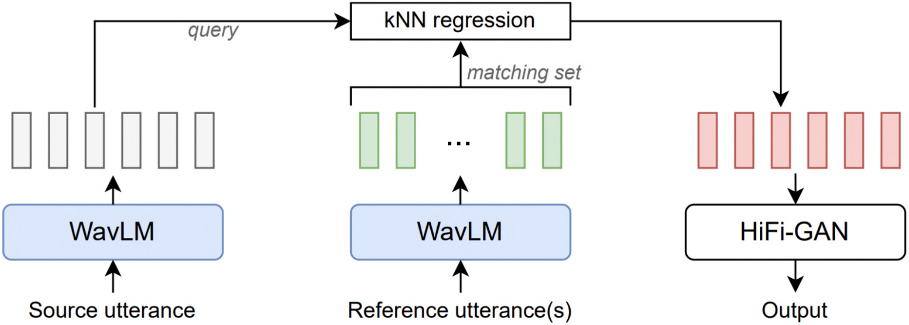
Whispered music conversion:

Source: Play Reference: Play Output: Play

Human-to-animal conversion:

Source: Play Reference: Play Output: Play

Voice conversion with stuttered reference speech



Source: Play Reference: Play Output: Play Baseline: Play (TTS)

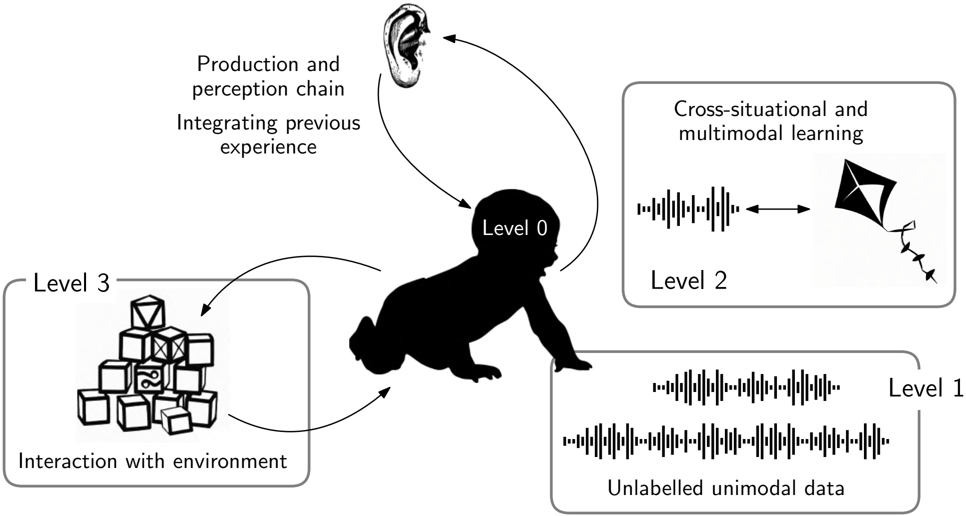
Source: Play Reference: Play Output: Play Baseline: Play (manual)

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



Production and perception chain

Integrating previous experience

Level 0

Cross-situational and multimodal learning

Level 2

Level 3

Interaction with environment

Level 1

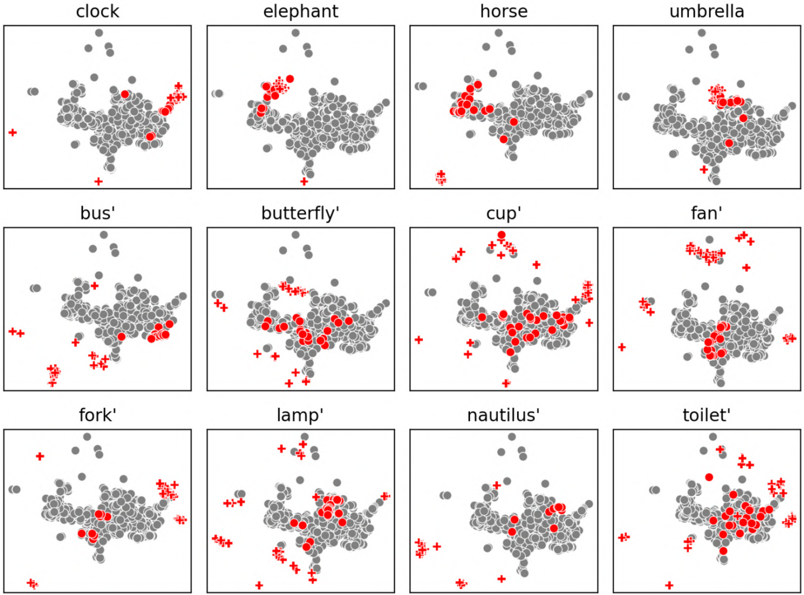
Unlabelled unimodal data

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

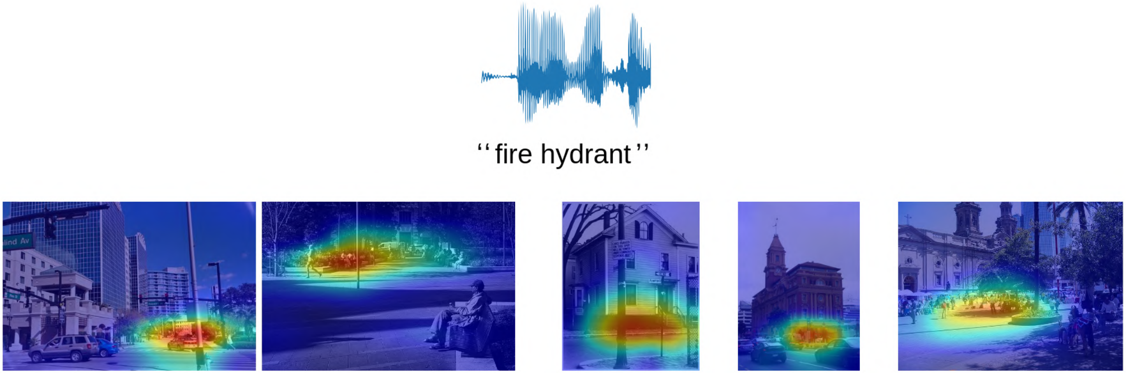
https://www.kamperh.com/

Mutual exclusivity results

		Model initialisation		Accuracy (%)	
		Audio (CPC)	Vision (AlexNet)	Familiar–familiar	Familiar–novel
1	Random baseline	N/A	N/A	50.19	49.92
2	MattNet	✗	✗	72.86	57.29
3		✗	✓	85.89	59.32
4		✓	✗	75.78	55.92
5		✓	✓	83.20	60.27


clock
elephant
horse
umbrella
bus'
butterfly'
cup'
fan'
fork'
lamp'
nautilus'
toilet'

Attention visualisation



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