Speech systems that emulate language acquisition in humans

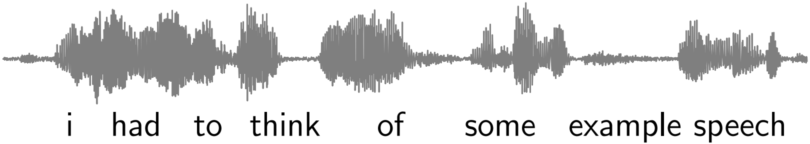
Swiss Data Science Centre, EPFL, Sep. 2023

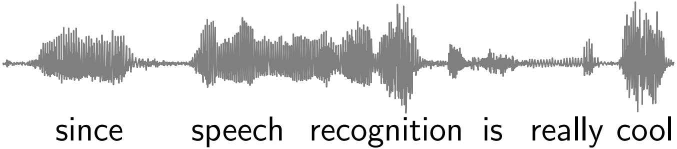
Herman Kamper

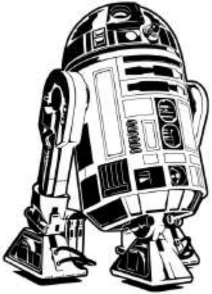
E&E Engineering, Stellenbosch University, South Africa

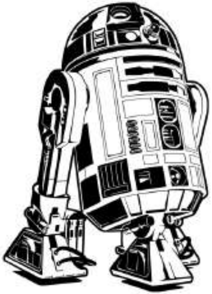
http://www.kamperh.com/

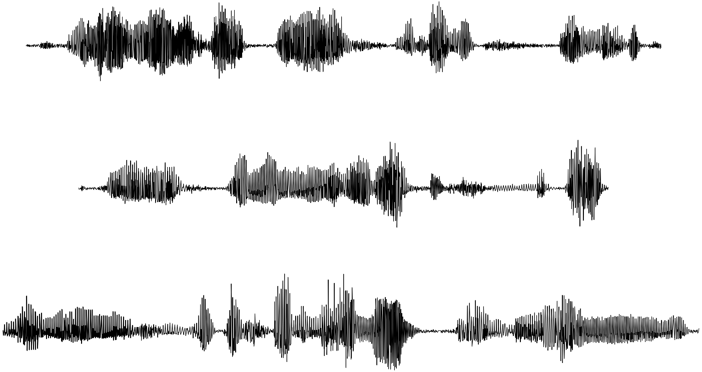
Supervised speech recognition and synthesis

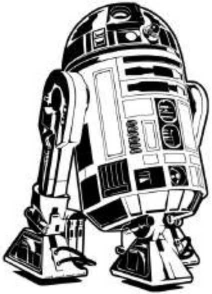


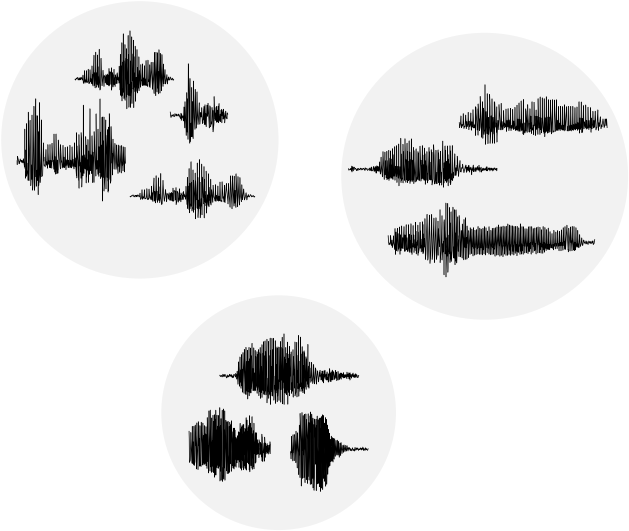


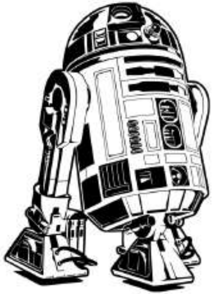


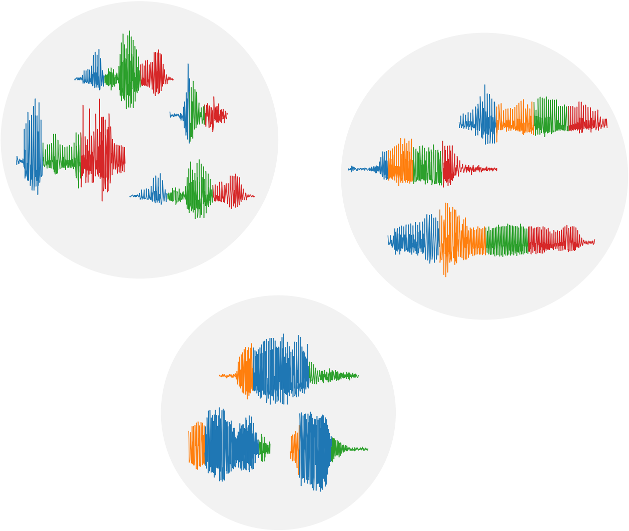












Why attempt to emulate language acquisition?

- Improvements in speech technology
- New insights and approaches for machines that learn
- New insights into human learning

This talk: Science and engineering

1. Cognitive models of language acquisition
2. Enabling new speech technology

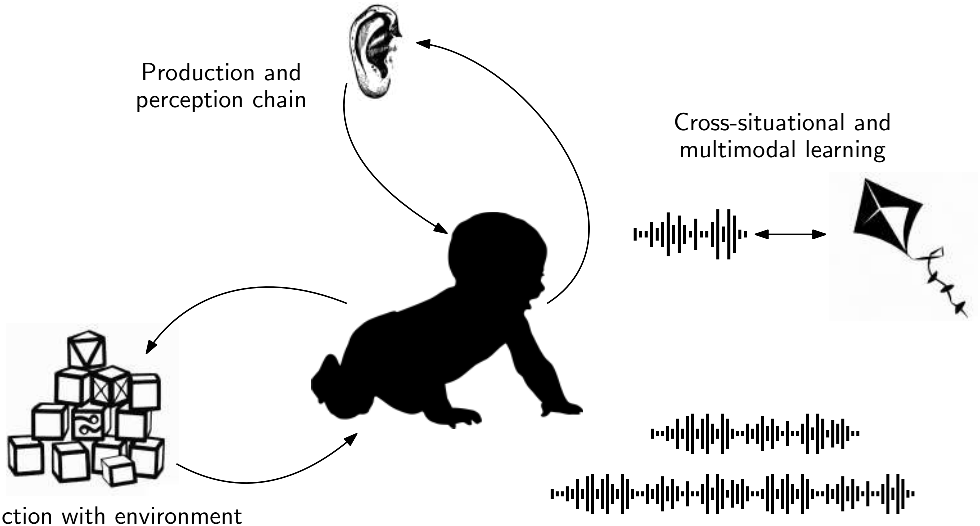
1. Cognitive models of language acquisition

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.

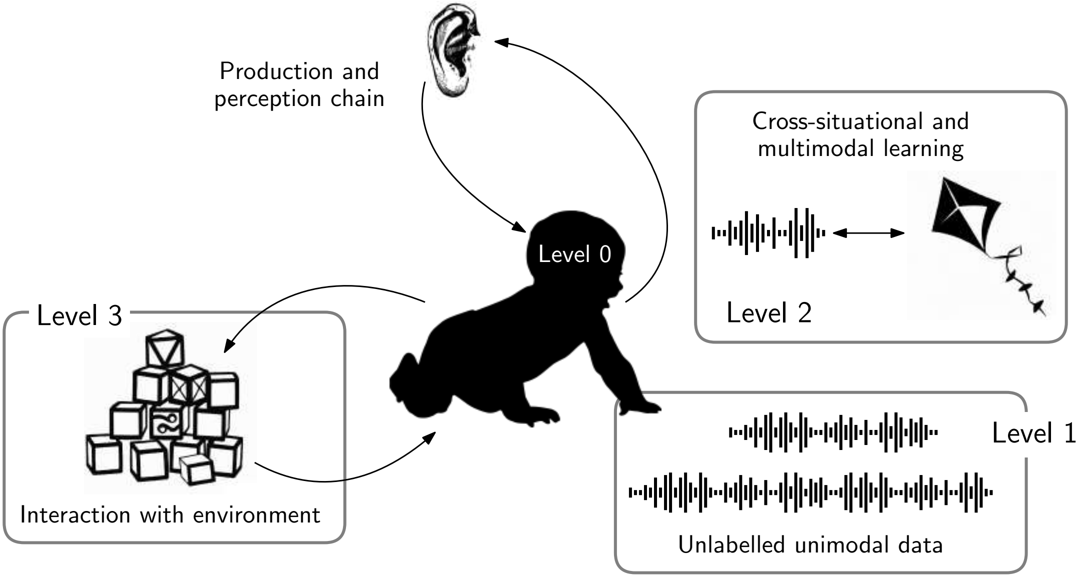


Production and perception chain

Cross-situational and multimodal learning

Interaction with environment

Unlabelled unimodal data



Production and perception chain

Level 0

Cross-situational and multimodal learning

Level 2

Level 3

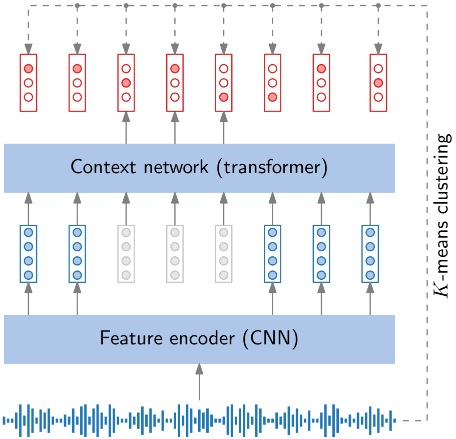
Interaction with environment

Level 1

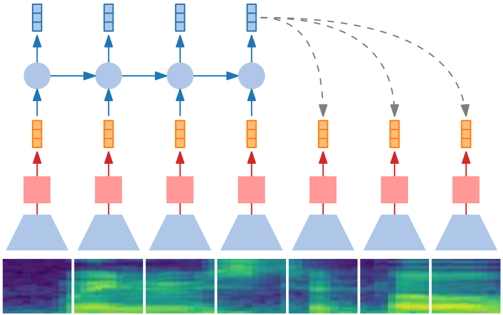
Unlabelled unimodal data

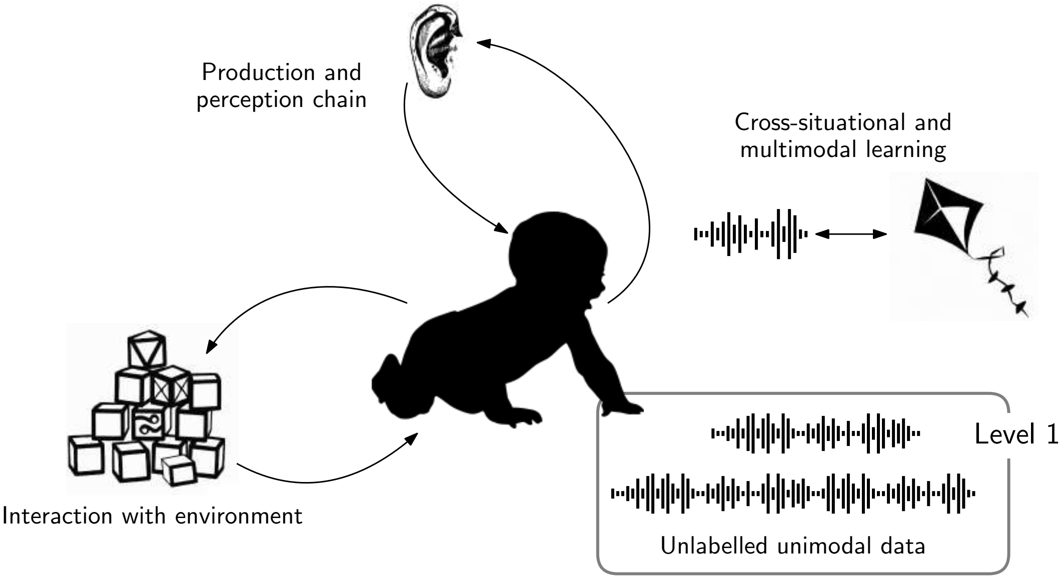
Large self-supervised spoken language models

HuBERT / WavLM:



Contrastive predictive coding (CPC):





Production and perception chain

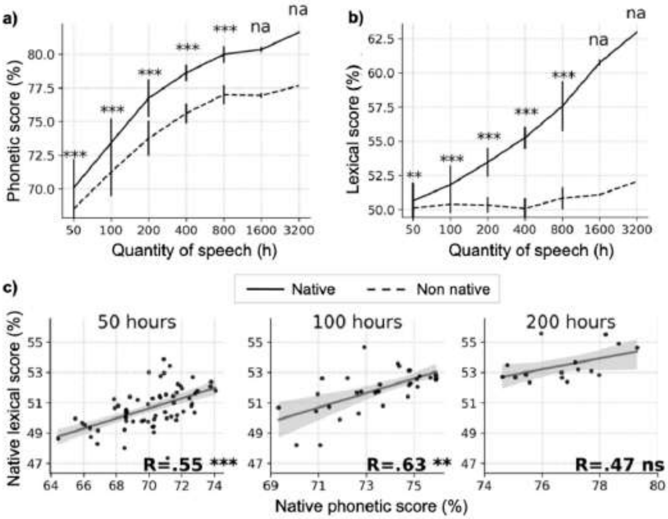
Cross-situational and multimodal learning

Interaction with environment

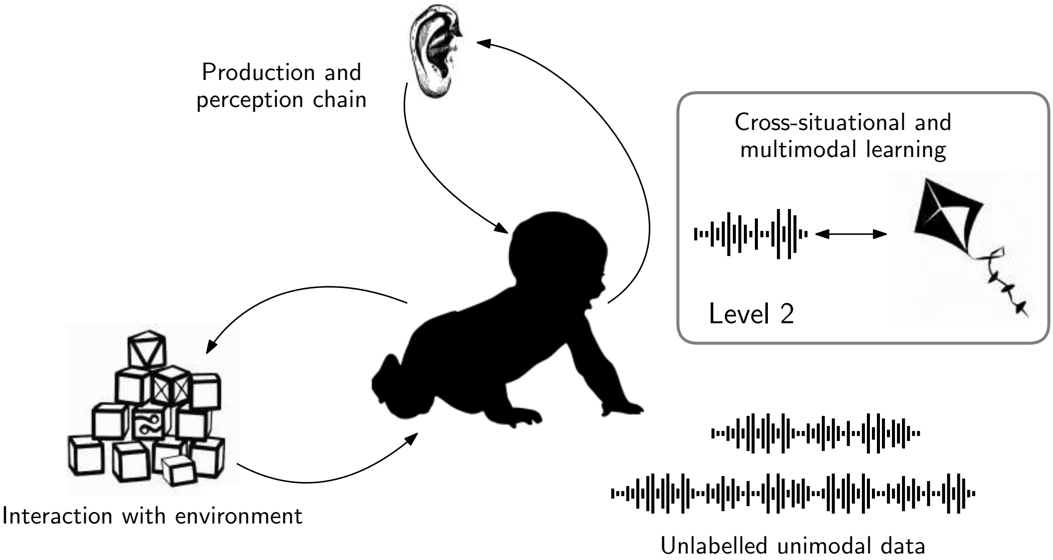
Level 1

Unlabelled unimodal data

Contrastive predictive coding as a language learner



M. Lavechin et al., "Can statistical learning bootstrap early language acquisition? A modeling investigation," *PsyArXiv*, 2022.



Production and perception chain

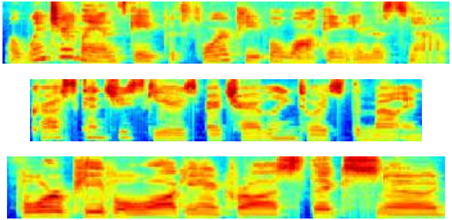
Cross-situational and multimodal learning

Level 2

Interaction with environment

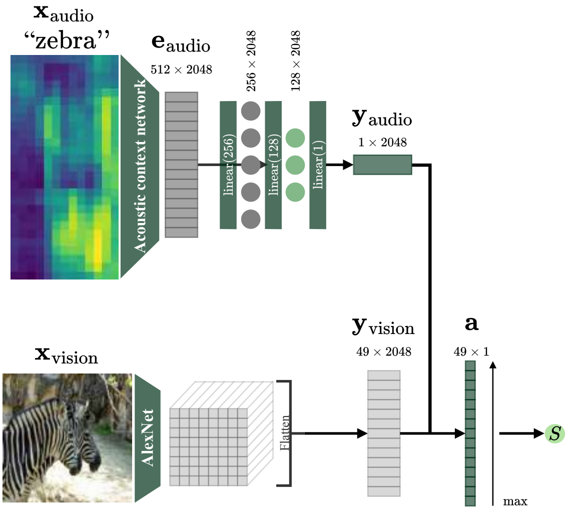
Unlabelled unimodal data

Using images for grounding speech



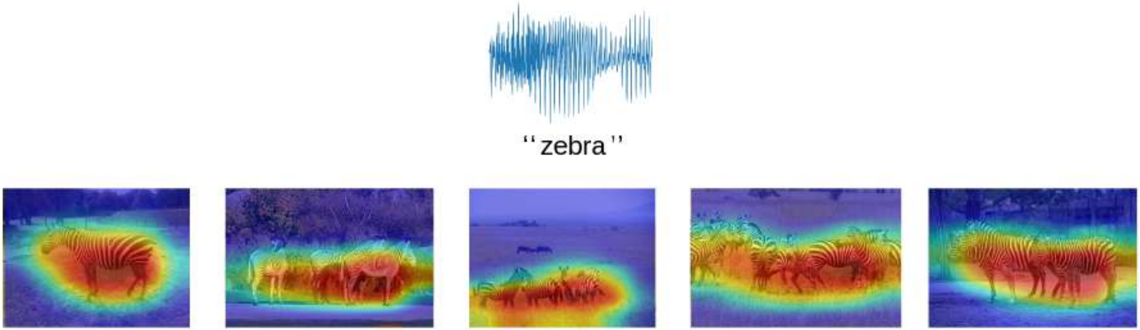
Harwath et al., "Unsupervised learning of spoken language with visual context," in *NeurIPS*, 2016.

Multimodal attention network (MattNet)

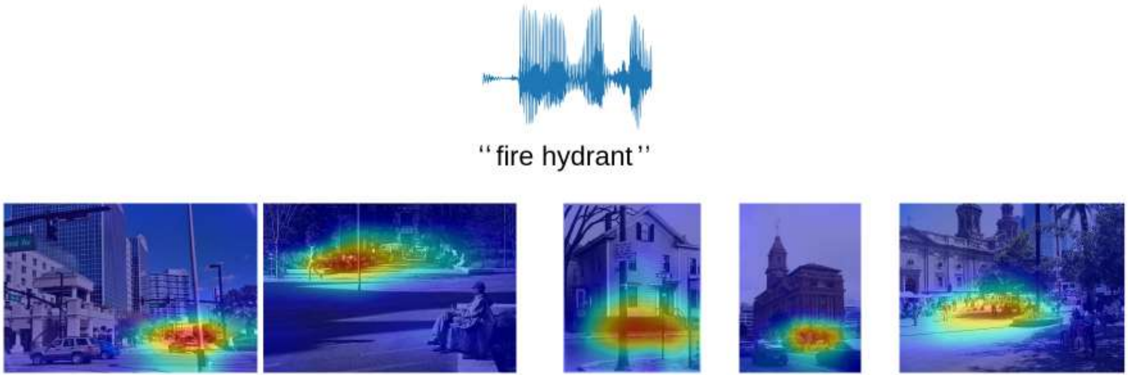


The acoustic context network is a CPC model trained on Places and LibriSpeech (level 1).

Attention visualisation



Attention visualisation



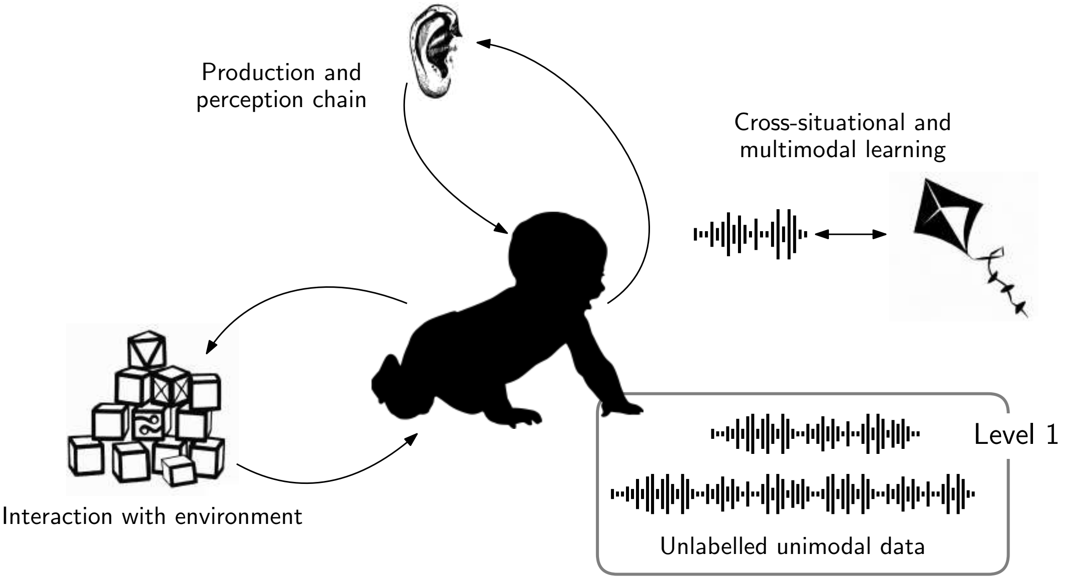
2. Enabling new speech technology: Voice conversion

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.



Production and perception chain

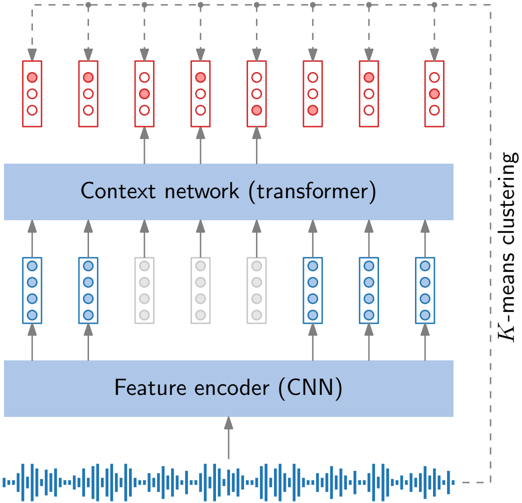
Cross-situational and multimodal learning

Interaction with environment

Level 1

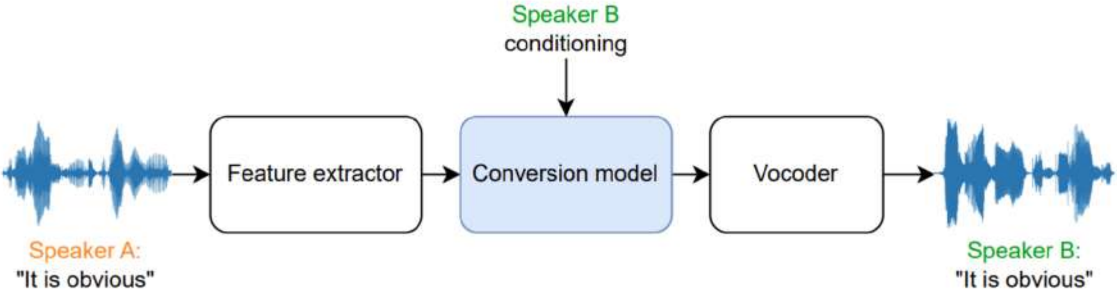
Unlabelled unimodal data

Large self-supervised spoken language models



HuBERT / WavLM

Voice conversion

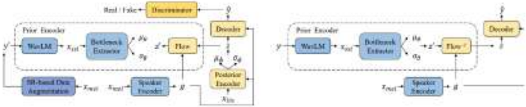


Source: Play

Reference: Play

Output: Play

Existing voice conversion systems



FreeVC [2022]

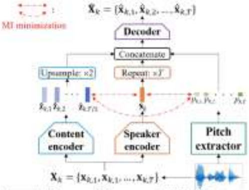
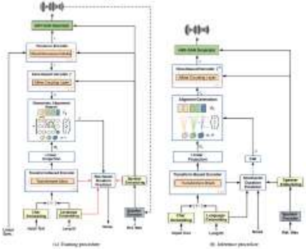


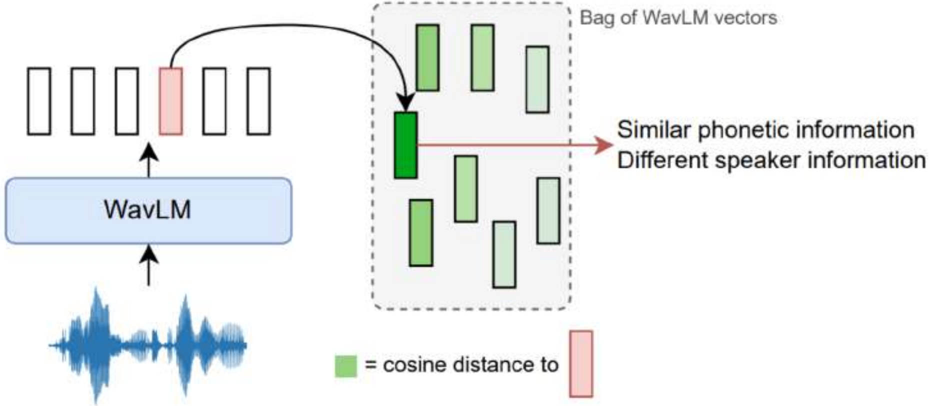
Figure 1: *Diagram of the proposed VQMIVC system.*

VQMIVC [2021]



YourTTS [2023]

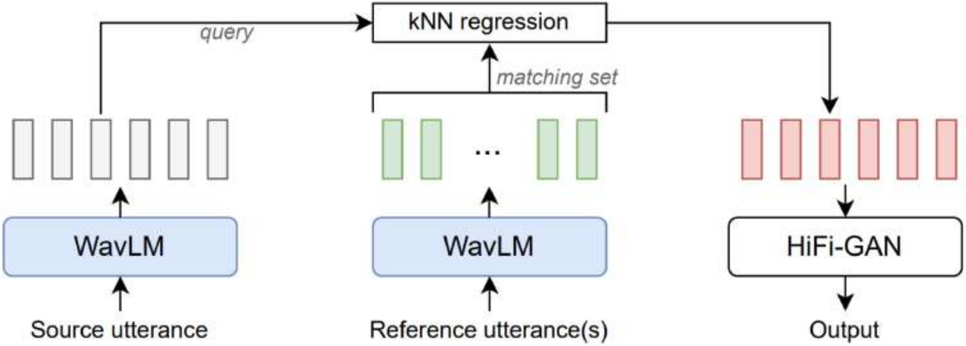
Our key idea



Similar phonetic information
Different speaker information

🟩 = cosine distance to 🟥

k-nearest neighbours voice conversion (kNN-VC)



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

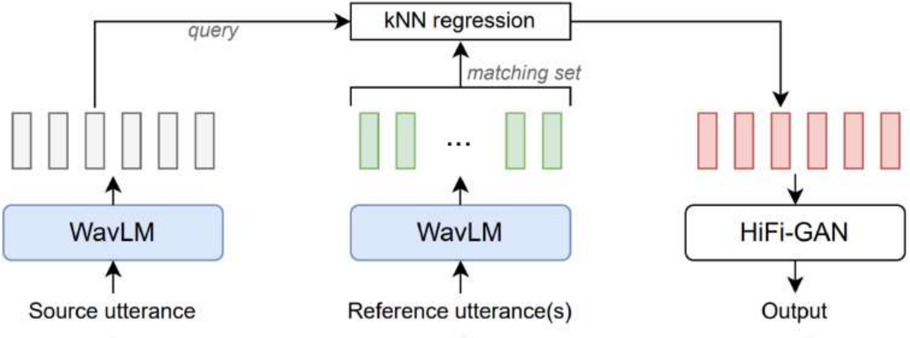
Whispered music conversion:

Source: Play Reference: Play Output: Play

Human-to-animal conversion:

Source: Play Reference: Play Output: Play

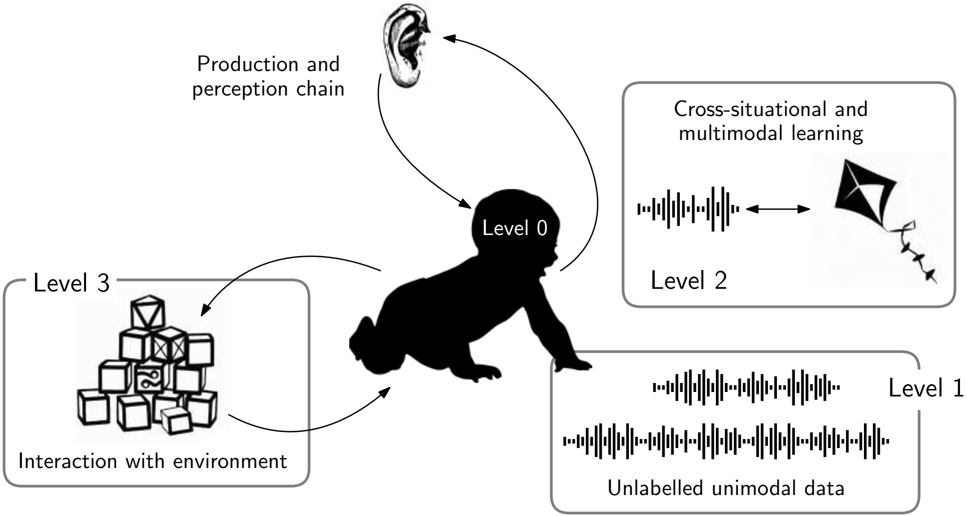
Voice conversion with stuttered reference speech



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

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

Conclusion



Production and perception chain

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