KONE FOUNDATION

ACADEMY OF FINLAND

Tampere University

Speech and Cognition

Speech and Cognition research group

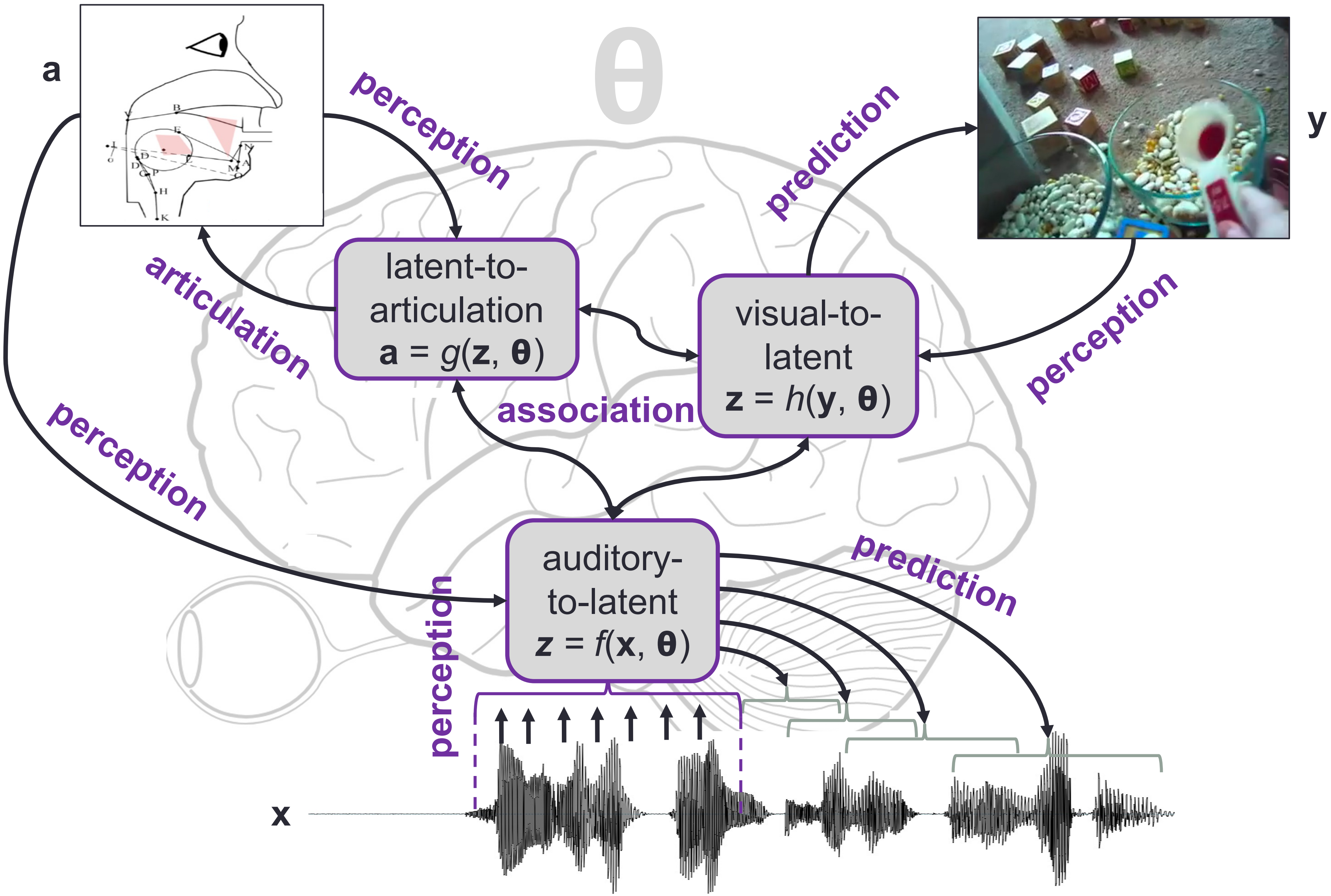
COMPUTATIONAL MODELING OF INFANT LANGUAGE LEARNING FROM REALISTIC-SCALE SPEECH AND AUDIOVISUAL INPUT

Khazar Khorrami & **Okko Räsänen**

Speech and Cognition research group
Signal Processing Research Center
Tampere University, Finland

Symposium on Computational Approaches to Early Language Development
ICIS-2024, Glasgow, UK

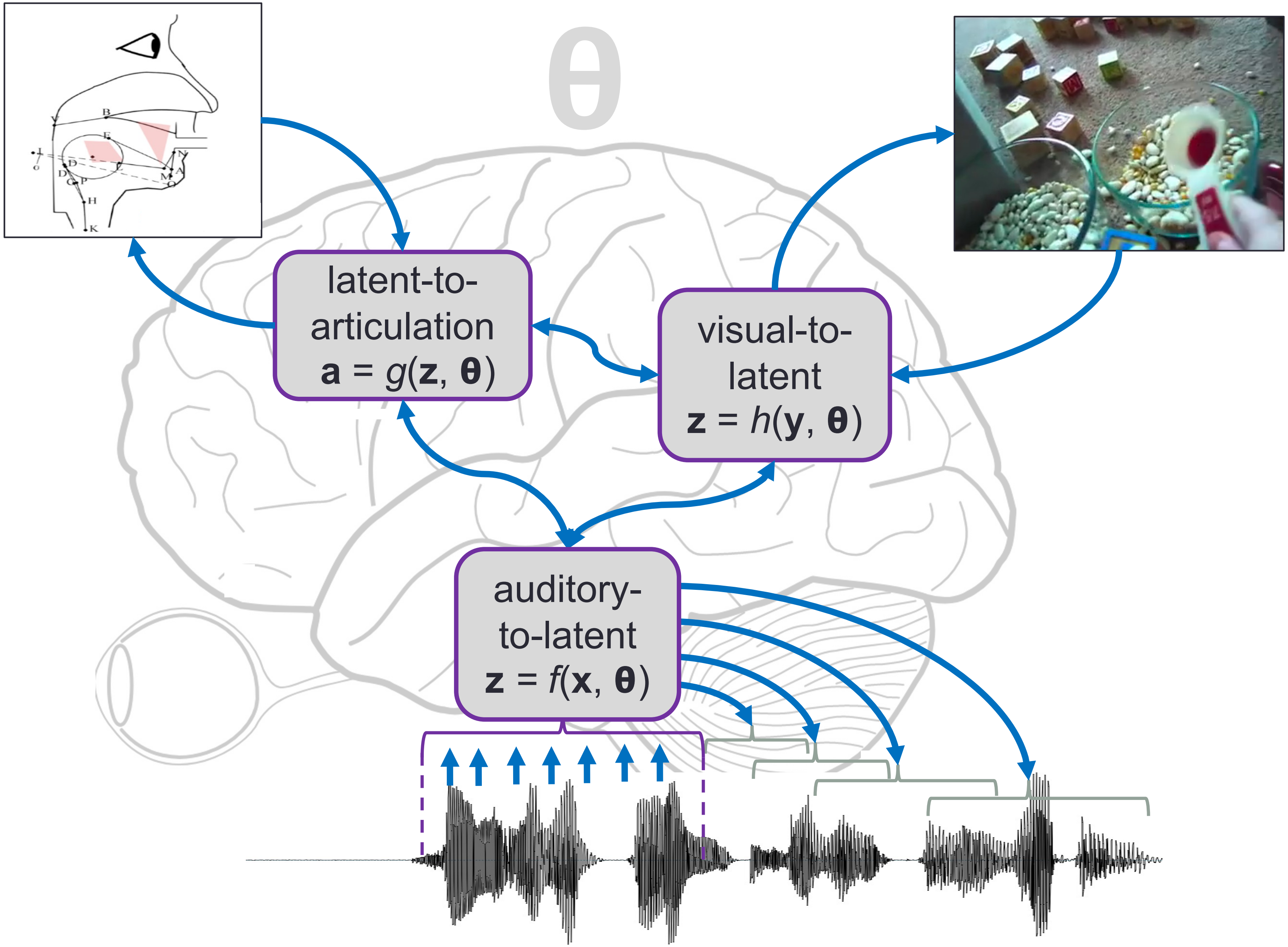
Photo from SEEDLingS dataset / Bergelson



Basic framework for computational modeling: statistical learning via predictive processing

(e.g., Rao & Ballard, 1999; Friston, 2010; Clark, 2013)

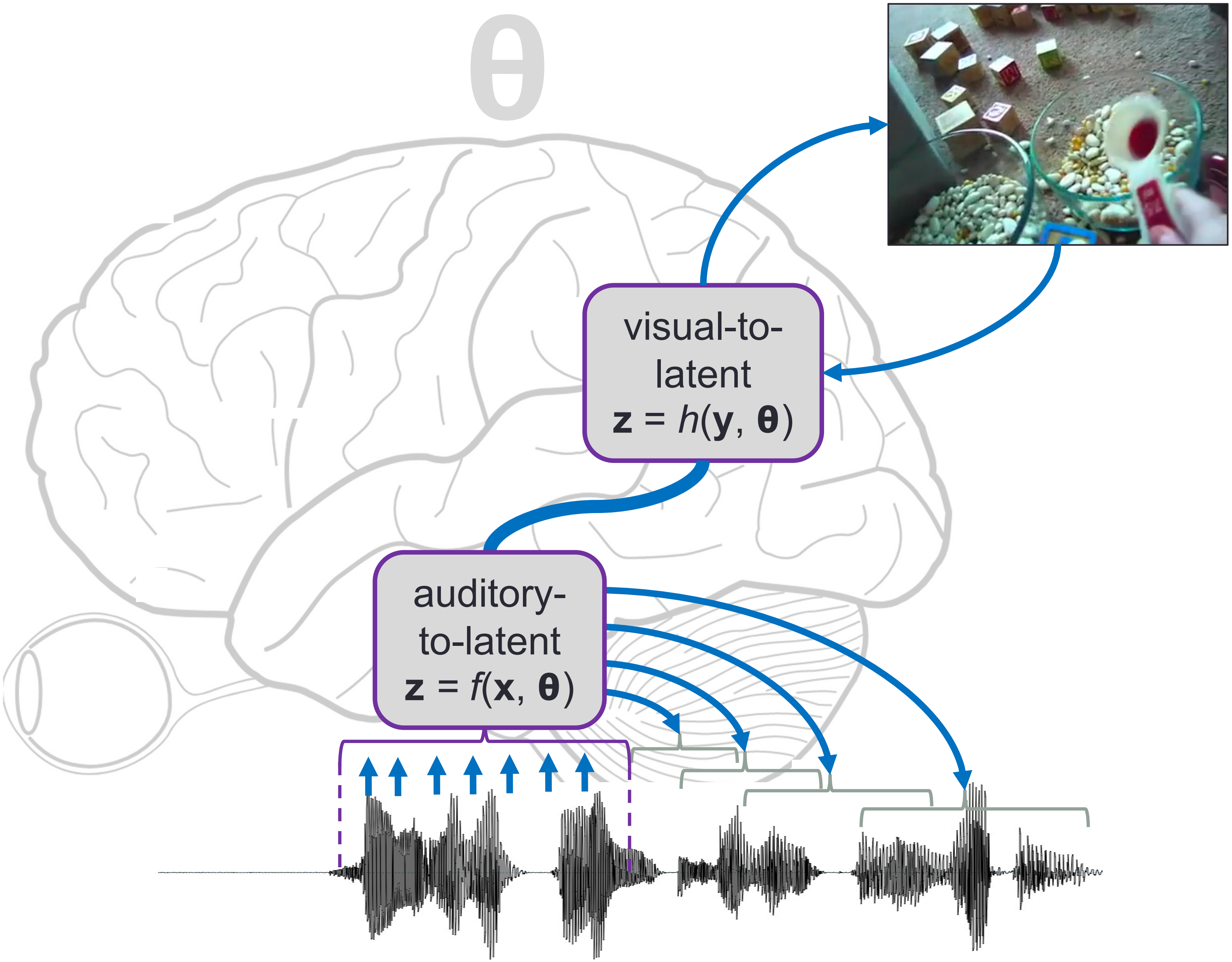
Photo from SEEDLingS dataset / Bergelson



Basic framework for computational modeling: statistical learning via predictive processing

(e.g., Rao & Ballard, 1999; Friston, 2010; Clark, 2013)

Photo from SEEDLingS dataset / Bergelson



This work: auditory + audiovisual statistical learning

Background & research question

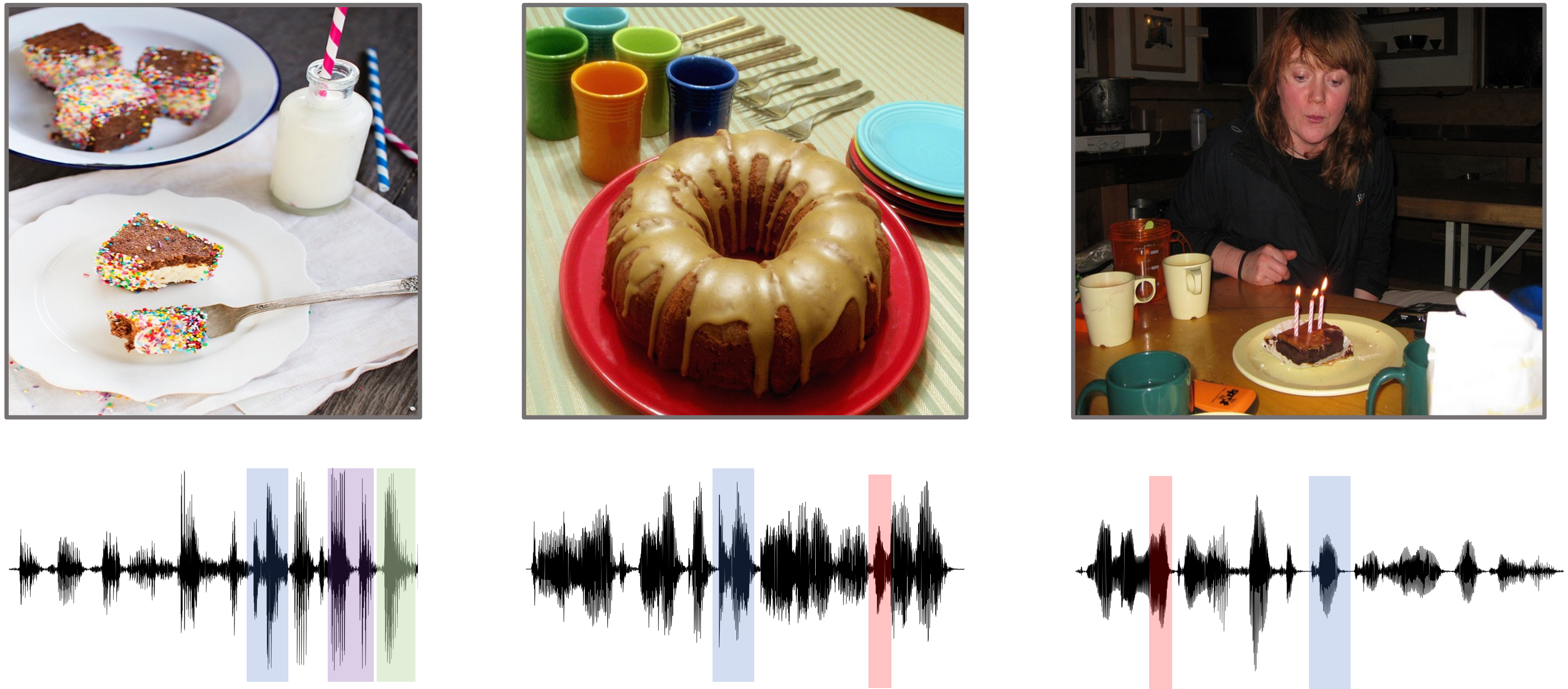
Previous research with audiovisual computational models:

- Learning from photographs and their spoken descriptions
  (e.g., Harwath & Glass, 2017; Alishahi et al., 2017; Merkx et al., 2019; Khorrami & Räsänen, 2021; Peng et al., 2023)
- Learning from infant head-mounted camera data + transcribed speech
  (Vong et al., 2024)
- Main findings: latent representations for phonemes, syllables, and words emerge as a side-product of audiovisual predictive optimization. ***No need for linguistic priors or proximal learning goals***!
- Limitations: models trained on thousands of speech-image pairs (“naming events”) or with simplified speech representations (text).

→ **Unclear if word learning succeeds from infant-scale sensory input with real speech**

→ **This work: simulate auditory and audiovisual learning with realistic-scale input.**

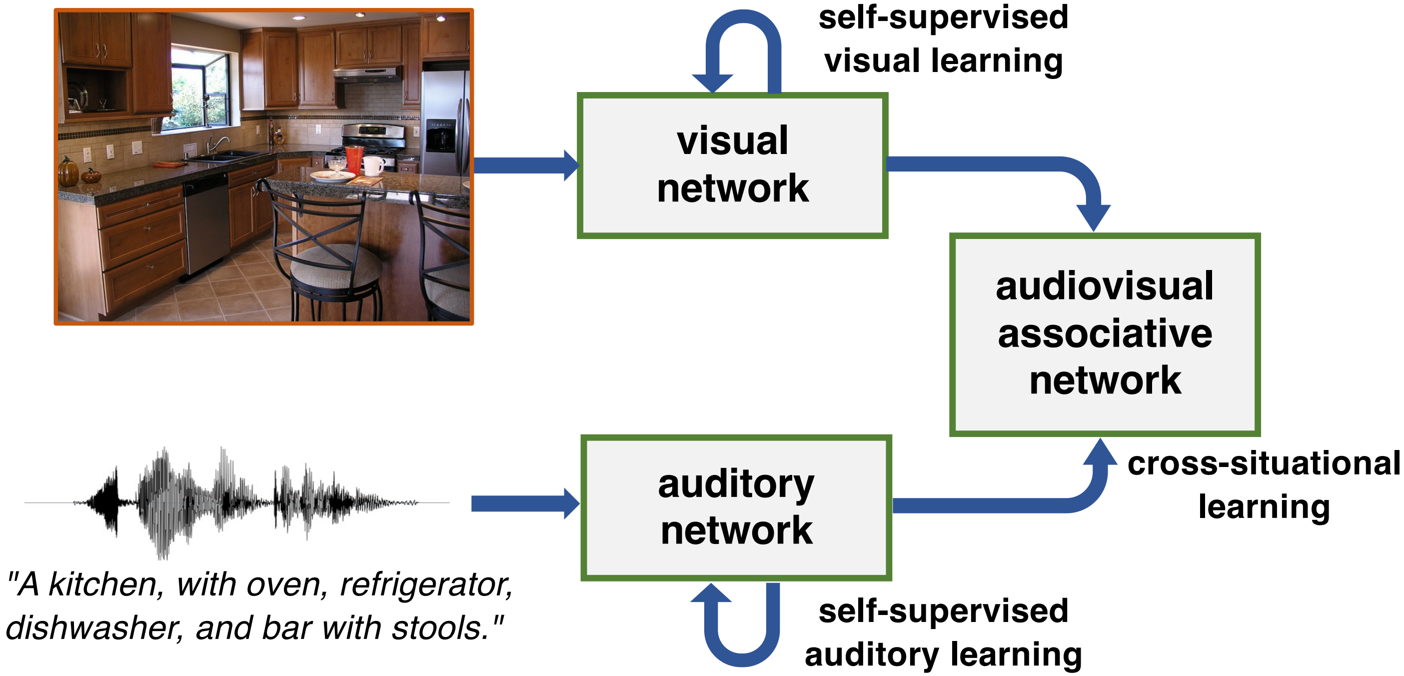
Referential ambiguity in audiovisual learning



Basic challenges:

- Segmentation problem in the auditory and visual domains ("where")
- Recognition problem in both domains ("what")
- Referential ambiguity across domains (e.g., Quine, 1960; Smith & Yu, 2008)

Model architecture (adapted from Peng & Harwath, 2022)

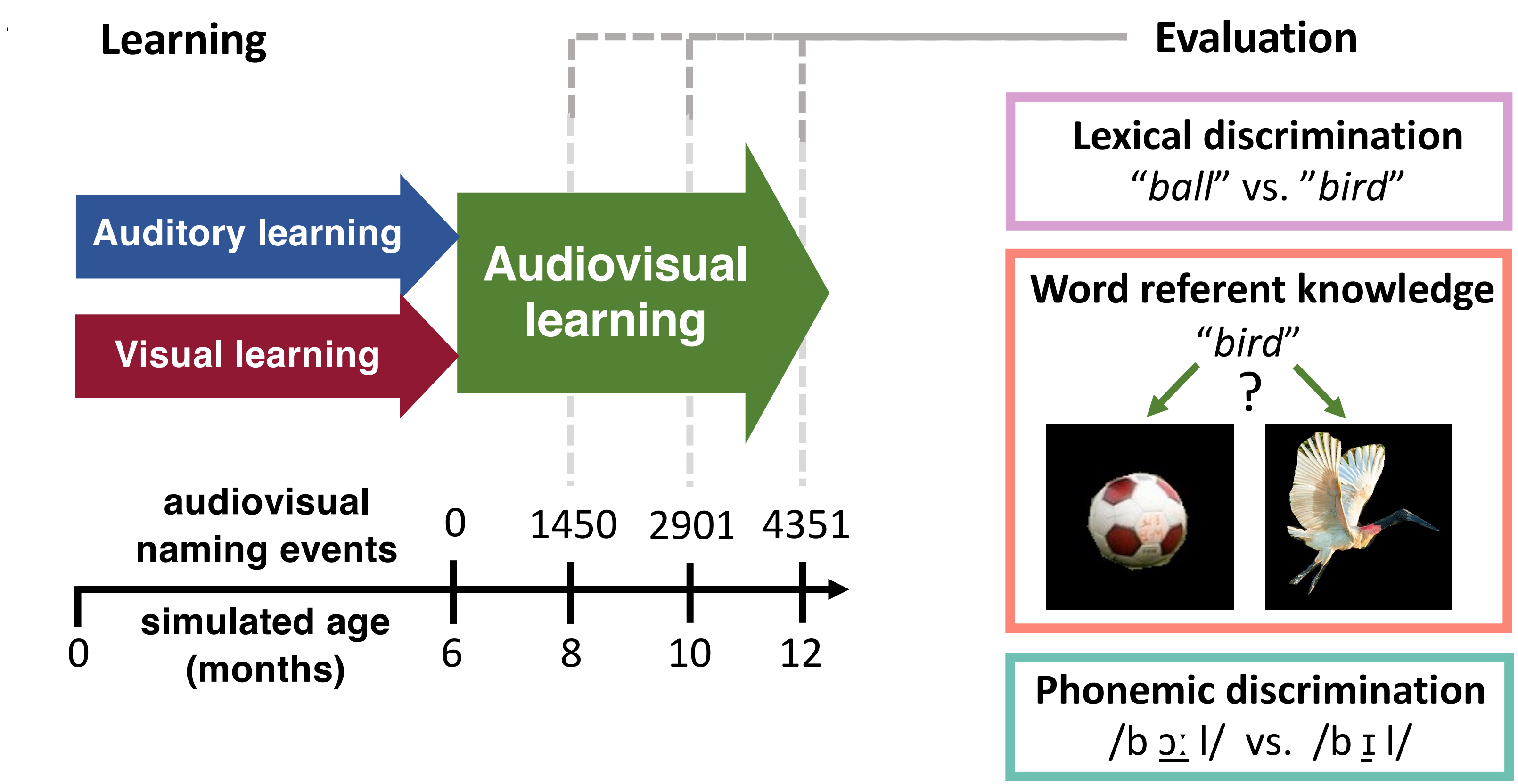


A deep neural network with three parts:

1) Visual encoder: DINO
2) Auditory encoder: Wav2Vec 2.0
3) Associative network: contrastive learning.

No supervision or data labels.
Only self-supervised (statistical) learning from sensory input.

Experimental setup



Training data design

Auditory learning:

1049 h of speech input to simulate 6 months of auditory learning (e.g., Cruz-Blandon et al., 2023; Coffey et al., 2024).

- From Librispeech + SpokenCOCO corpora

Audiovisual learning:

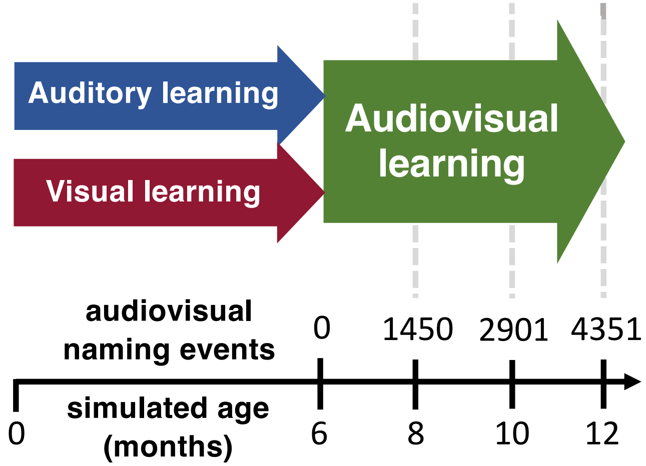
Photographs and their spoken descriptions from SpokenCOCO dataset.

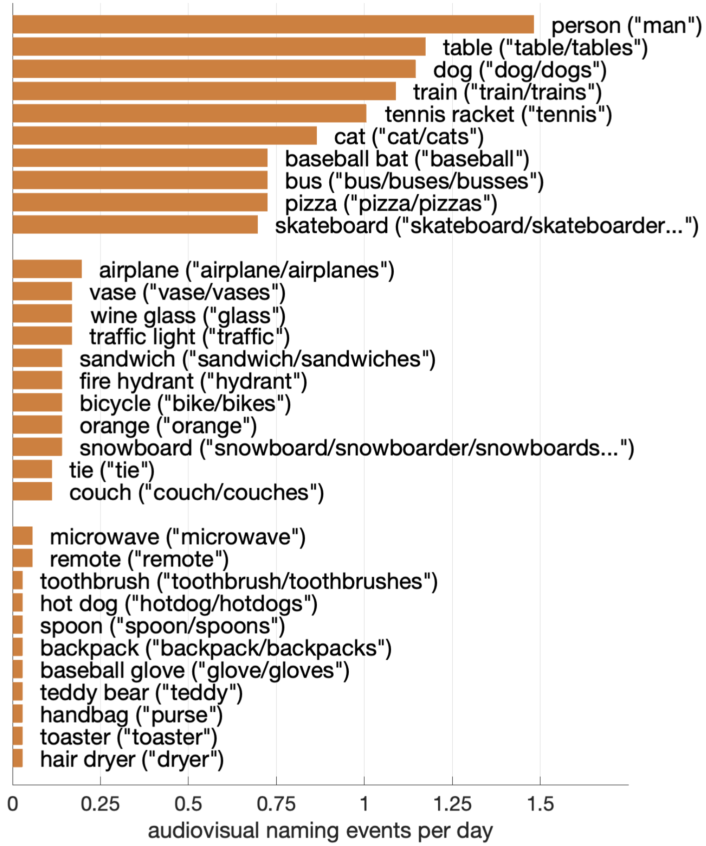
Empirical estimates of daily object naming rates for the 80 most frequent word-object pairs (from Clerkin & Smith, 2019; 2022).

→ Extrapolate counts to 2, 4 or 6 months

→ Select images + utterances that satisfy the statistics.

- Words per utterance: 11.3 ± 2.59
- Content words per utterance: 5.87 ± 1.47
- Visual targets per image: 2.9 ± 1.84





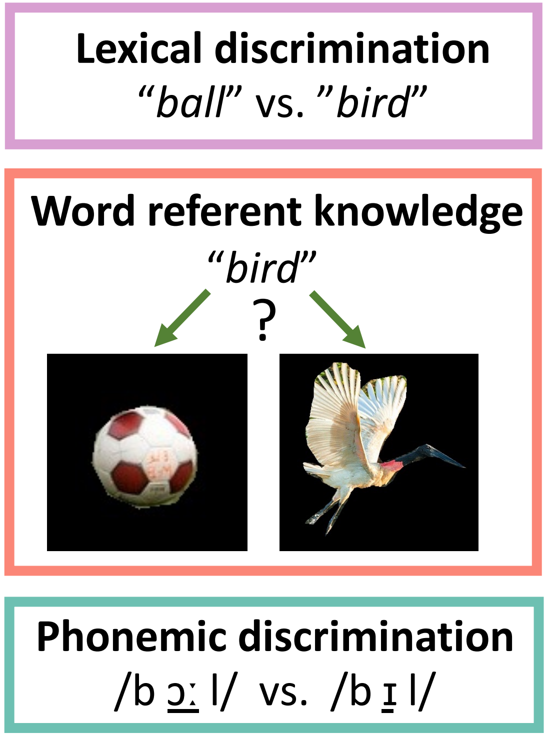
Model evaluation

Evaluate model at 6, 8, 10, and 12 months for:

- **Phonemic discrimination** (ABX-test; Schatz et al., 2023)

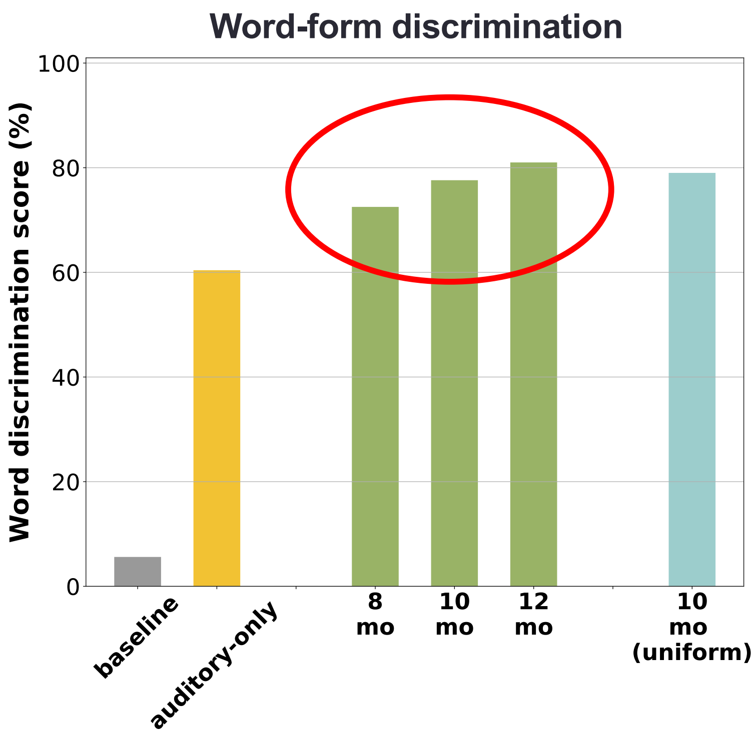
- **Auditory word-form discrimination** (CDI-Lextest, Khorrami et al., 2023).

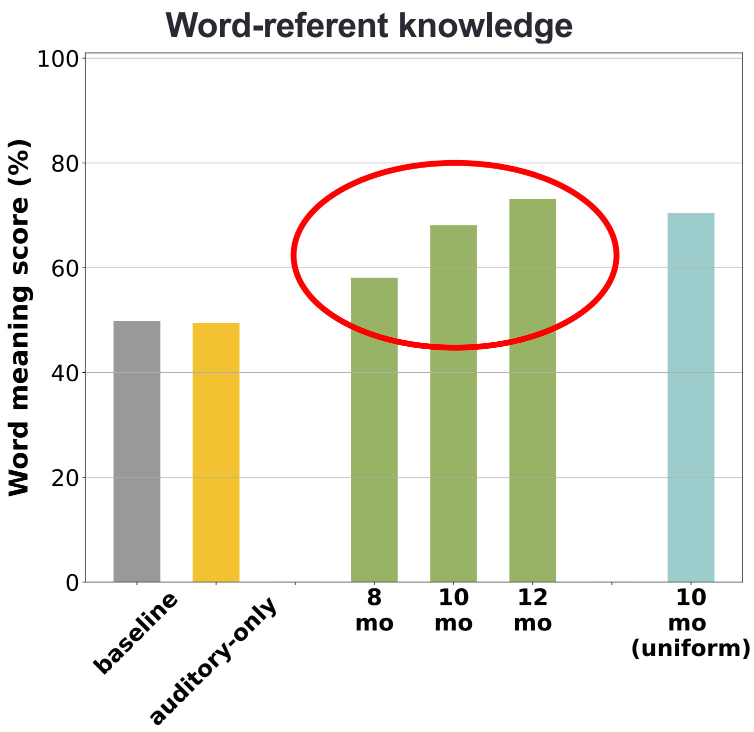
- **Word referent knowledge** for the 80 audiovisual concepts in SpokenCOCO (an audiovisual forced-choice task)



Results

Word referent learning didn't work without the auditory learning stage.

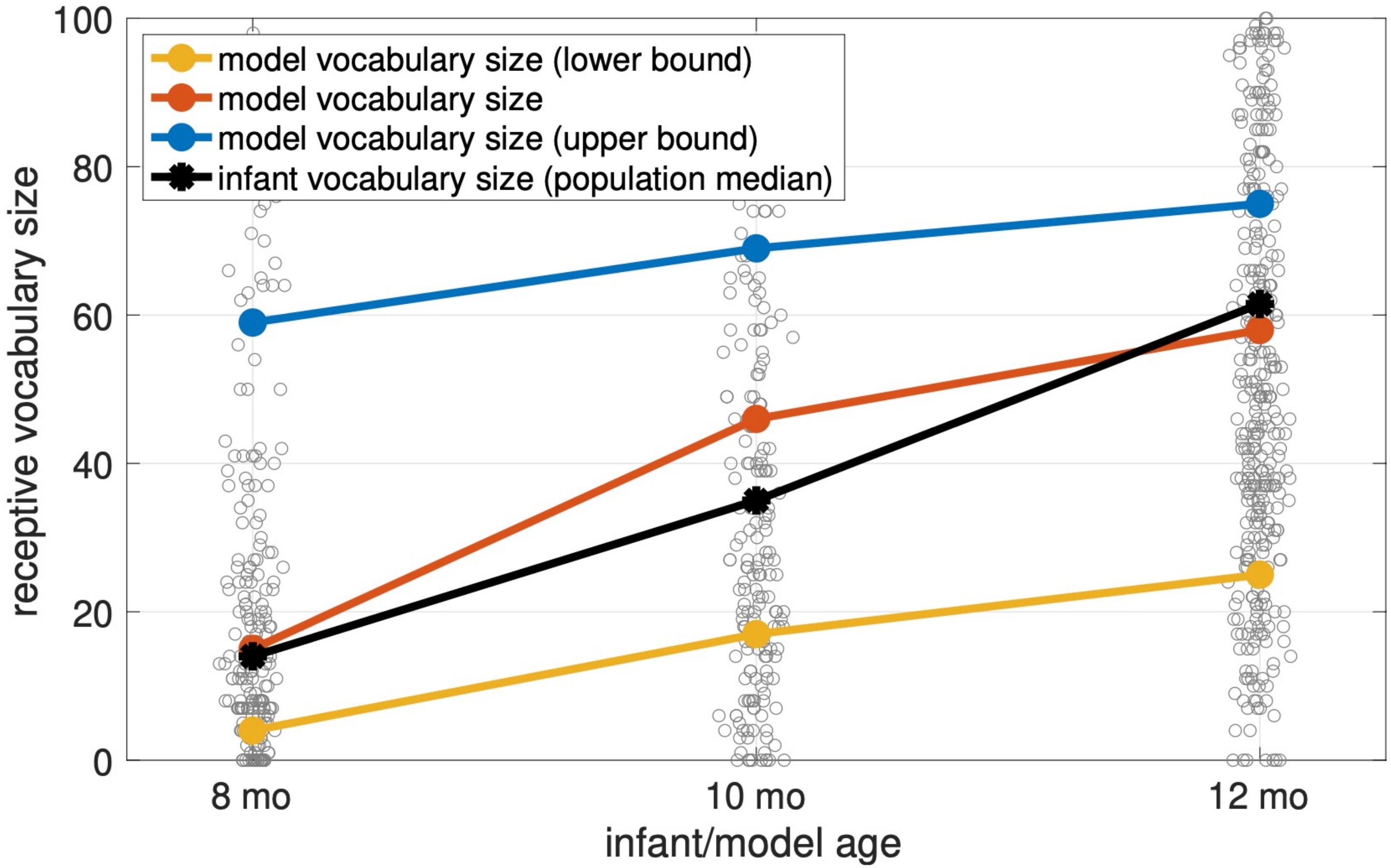




Phoneme error rate: 7.1% after 6 mo auditory learning (chance: 50%). No change during audiovisual stage.

Phoneme and word comprehension skills emerge from plausible-scale data!

Vocabulary growth: model vs. CDI-norms



CDI data: North-American infants, receptive lexicon (from Wordbank; Frank et al., 2017)

Conclusions

The model succeeds in learning proto-lexical (and phonemic) representations from infant-scale input.

Learning operates on *real speech* and images, and *without linguistic priors*, data labels, or other strong constraints.

Supports the idea of statistical learning as a means to boostrap early language acquisition.

Supports the “Latent Language Hypothesis”, according to which linguistic structures are not proximal targets of learning, but side products of predictive optimization (e.g., Khorrami & Räsänen, 2021, *Lang. Dev. Res*).

- No need to “cluster” phone(me)s or segment words as intermediate stages. Only prediction within and across sensory modalities.

Paper pre-print:
https://arxiv.org/abs/2406.05259

The end

Speech and Cognition

Speech and Cognition research group

khazar.khorrami@tuni.fi
okko.rasanen@tuni.fi

https://webpages.tuni.fi/specog/