EXPLORING THE ROLE OF DIFFERENT MODALITIES IN EARLY LANGUAGE ACQUISITION

Okko Räsänen

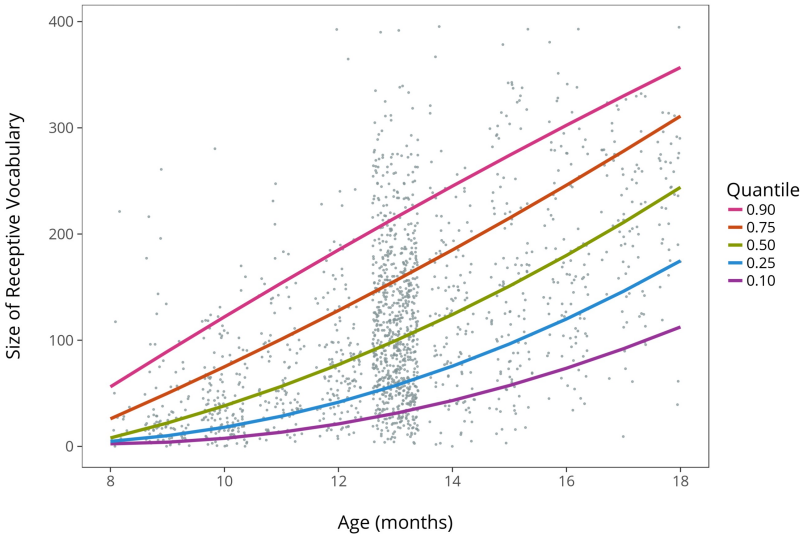
Faculty of Information Technology and Communication Sciences
Tampere University, Finland

29.4.2022

Infant language learning

Children learn their native language without explicit supervision.

- Interaction with environment and other agents sufficient.
- E.g., comprehension of first words around 8 mo, production at 12 mo.



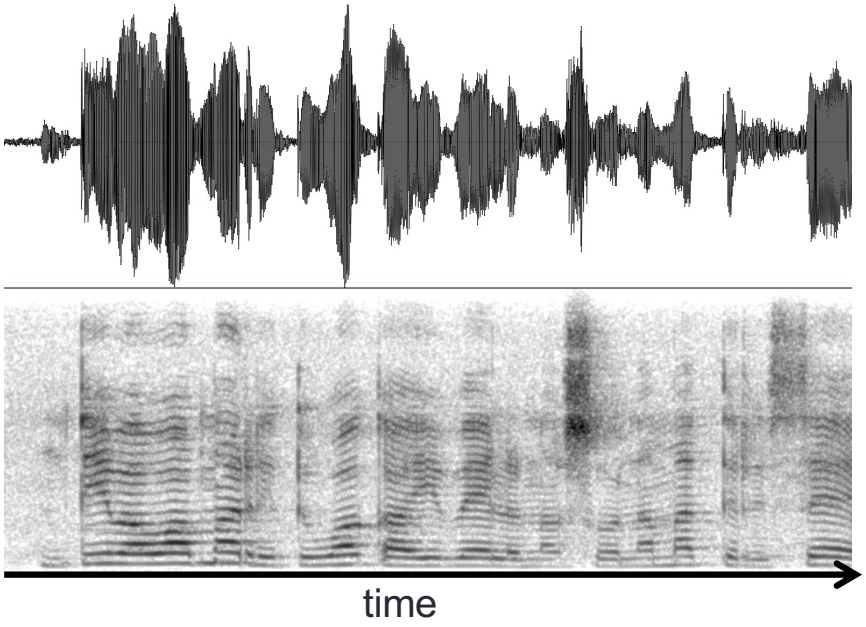
http://wordbank.stanford.edu

Research on child language development (CLD) tries to understand this process.

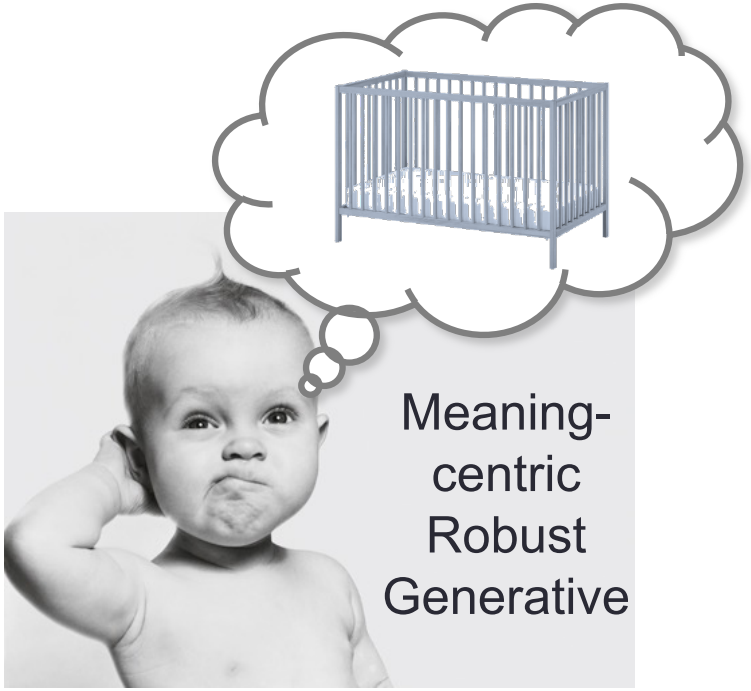
Some key questions:

- What are the learning mechanisms?
- What is the role of input (quality and quantity)?
- What are the underlying language representations?
- What is driving individual differences?

Infant language learning



Continuous
Noisy
Variable
Opaque
Lack of universal cues



"ntivovomarito wa kayivela naswona matirhise"

Discrete (categorical)
Invariant
Hierarchical
Compositional

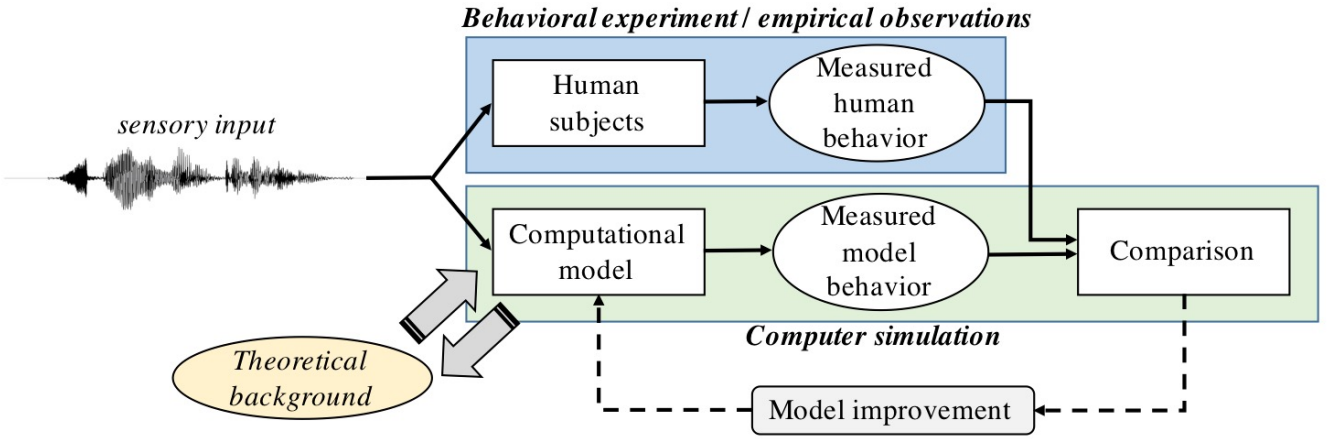
Despite decades of research, *we still lack a holistic model of the process.*

Computational modeling of CLD

Computational modeling as a means to study CLD.

The basic idea:

1. Implement models of infant learning as computational algorithms (~speech signal processing and machine learning)
2. Expose the models to sensory data similar to what infants have access to (at least speech).
3. Compare model behavior to human data.
4. Iterate to improve the model.



Computational modeling as statistical learning

Empirical CLD research: *language learning by capturing statistical regularities in the input* ("statistical/distributional learning").

- *Phonetic categories* from distributions of acoustic features (Kuhl et al., 1992; Maye et al., 2002).
- *Word segmentation* with transition probabilities of syllables (Saffran et al., 1996).
- *Word meanings* from co-occurrence probabilities of words and concurrent visual inputs (Smith & Yu, 2008)

Unsupervised and multimodal machine learning ~ *representation learning by capturing statistical regularities in the input*.

Computational modeling as statistical learning

What is common in different types of (human) statistical learning?

Uncertainty minimization, i.e., ***maximization of input predictability***.

- Cf. maximum-likelihood (ML) or maximum-a-posteriori (MAP) estimation.
- An alternative formulation: maximization of information gain.

”***Predictive brain hypothesis***” (e.g., Rao & Ballard, 1999; Friston, 2010; Clark, 2013).

Prediction and language learning?

Classical approach to CLD: study and model individual language phenomena separately.

- Model 1 for phonemic learning, model 2 for word segmentation etc..

For infants, no direct utility for solving the individual sub-tasks.

Where does the language come from?

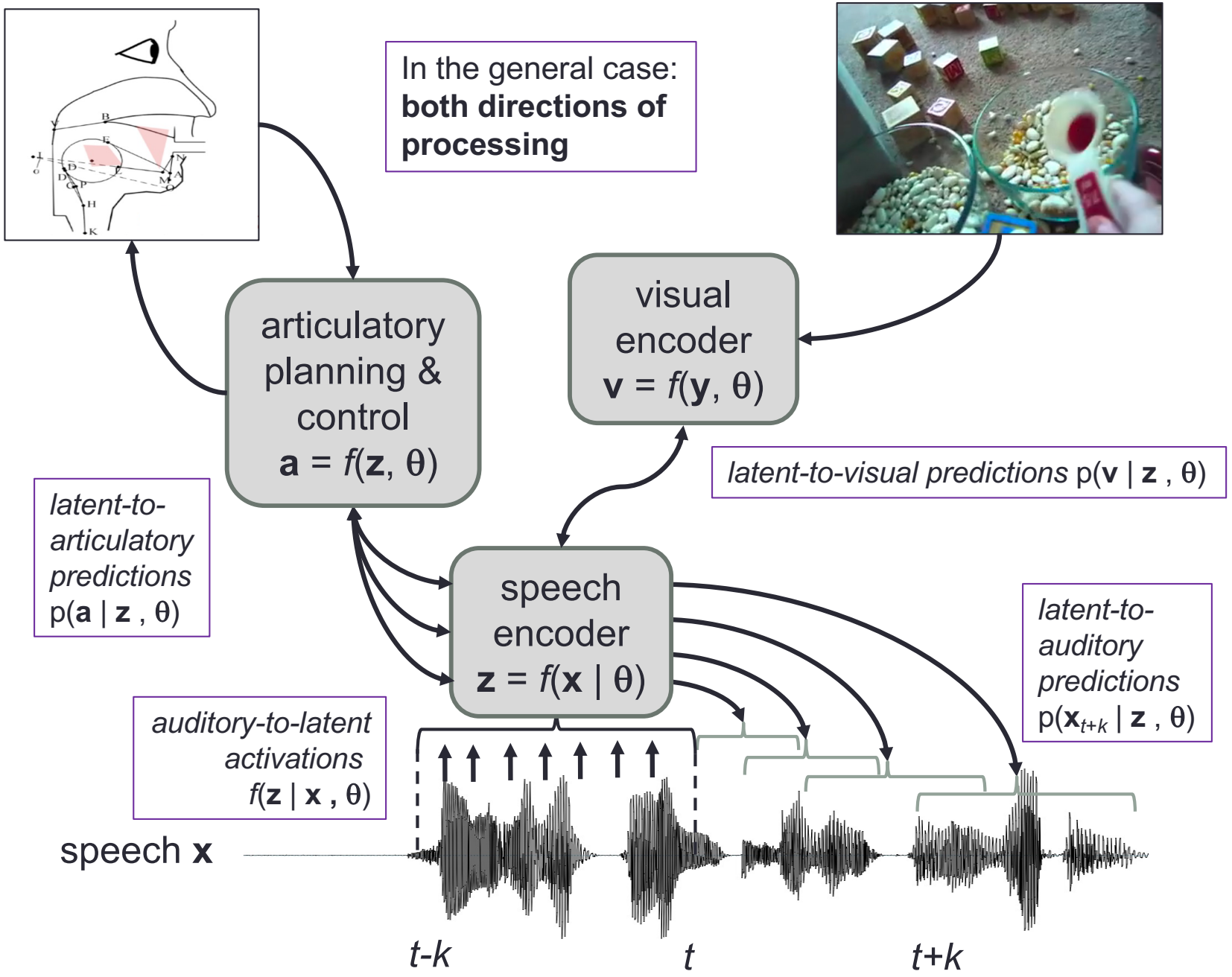
Latent language hypothesis (**LLH**; Khorrami & Räsänen, 2021; see also Räsänen & Rasilo, 2015):

Instead of being "proximal targets" of learning, ***language in the cognitive system emerges as latent representations that support predictive processing in the brain***.

Prediction and language learning?

If we want to test LLH, what are the *predictive tasks* that would facilitate emergence of language representations and skills?

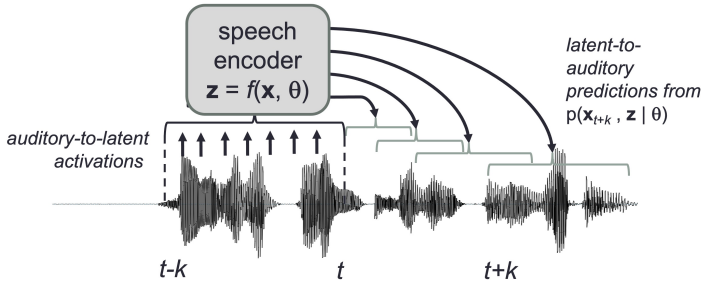
Depends on the available sensory inputs (and motor outputs)!


In the general case:
both directions of processing
articulatory planning & control
a = *f*(**z**, θ)
visual encoder
v = *f*(**y**, θ)
latent-to-visual predictions p(**v** | **z** , θ)
latent-to-articulatory predictions
p(**a** | **z** , θ)
speech encoder
z = *f*(**x** | θ)
latent-to-auditory predictions
p($\mathbf{x}_{t+k}$ | **z** , θ)
auditory-to-latent activations
f(**z** | **x** , θ)
speech **x**
t-k
t
t+k

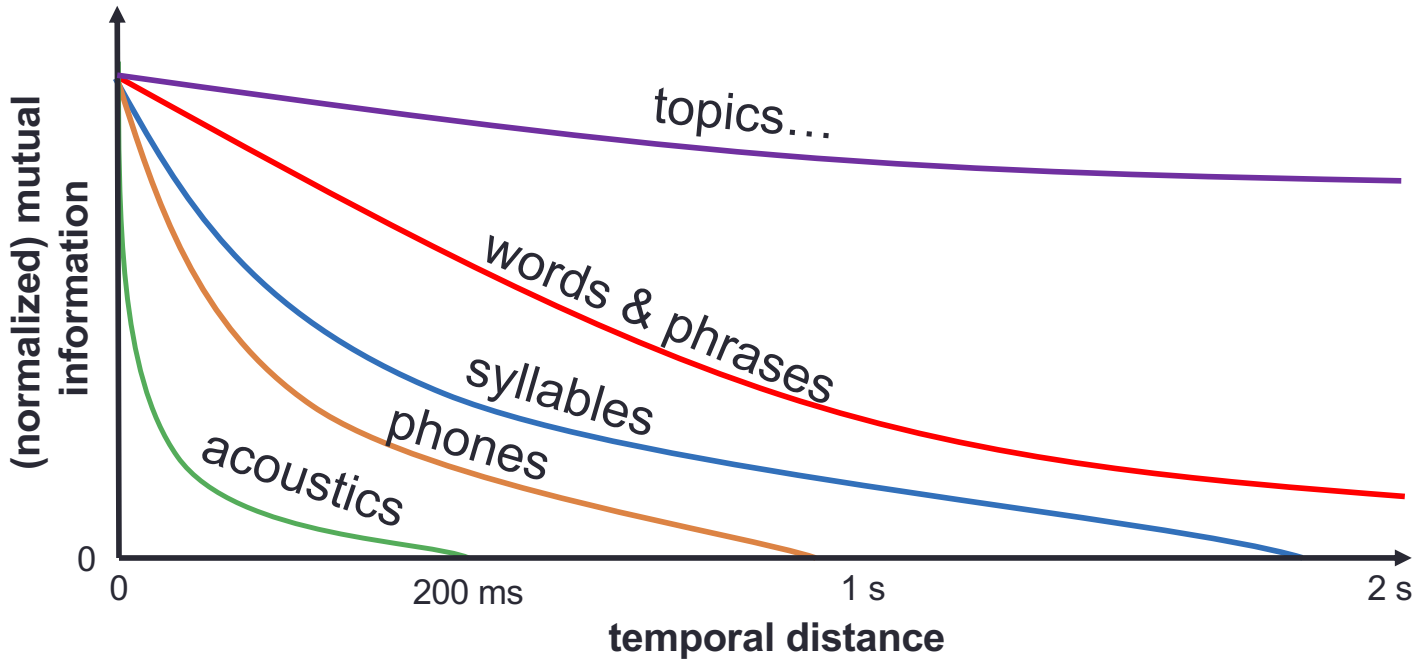
Language learning with auditory predictions

Speech-to-speech prediction

Speech-to-speech prediction
→ requires efficient coding of speech

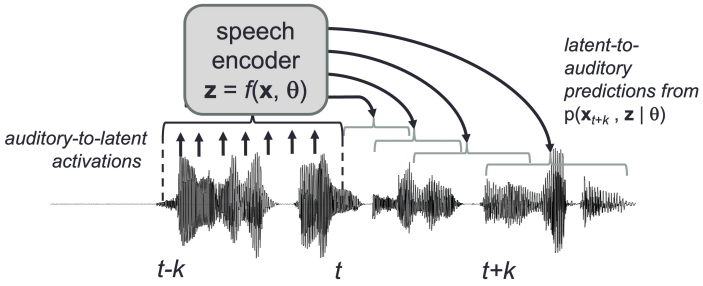


Increasing prediction distances require increasing levels of abstraction.



Speech-to-speech prediction

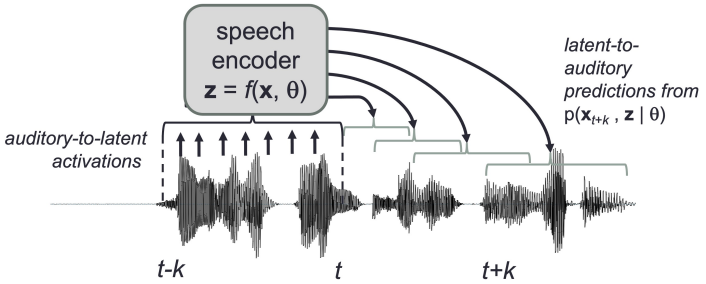
Speech-to-speech prediction
→ requires efficient coding of speech



Increasing prediction distances require increasing levels of abstraction.

Learning an efficient model for temporal prediction
→ *emergence of latent linguistic representations*?

Speech-to-speech prediction



Potential advantages:

- An “ideal linguistic unit” extractor (e.g., phonemes & words) as the optimal solution?

→ learning of distributional semantics similar to text-based systems (cf., LSA, word2vec, BERT, GPT-3 etc.)?

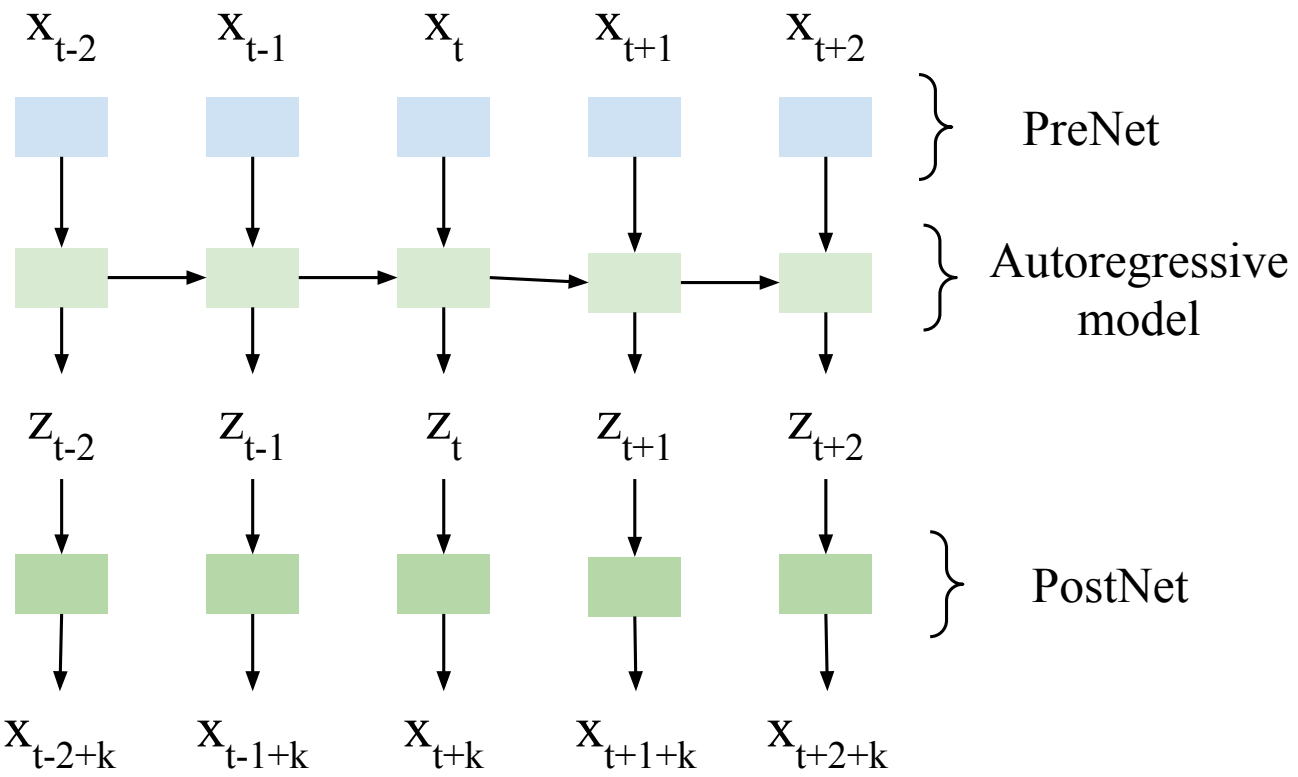
Potential limitations:

- Unimodal distributional semantics ***fail to explain all variance in human semantic data*** (e.g., De Deyne et al., 2021; Merkx et al., 2022).
- Speech-based representations ***must interface with other modalities and motor system*** to play any role in the cognitive system.

How much can speech-to-speech models ultimately learn?
(from the amount and type of input available to infants!)

Autoregressive predictive coding (APC)

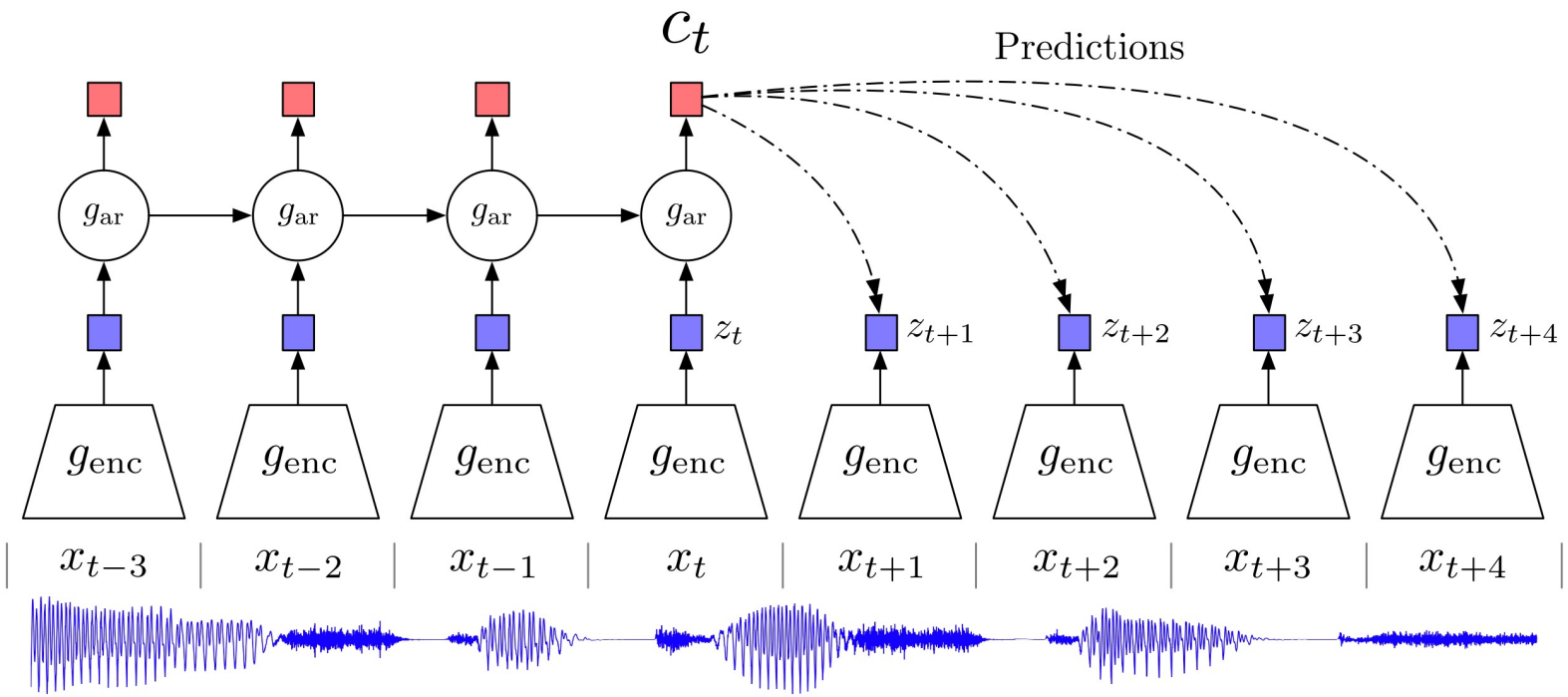
Autoregressive predictive coding (APC; Chung et al., 2019): predict future frames $\mathbf{x}_{t+k}$ of speech signal from $\mathbf{x}_0$, $\mathbf{x}_1$, …, $\mathbf{x}_t$



Contrastive predictive coding (CPC)

In APC, original signal features are predicted across time.

In **contrastive predictive coding** (CPC; van den Oord et al., 2018), the model predicts *its own latent vectors* extracted from the signal.

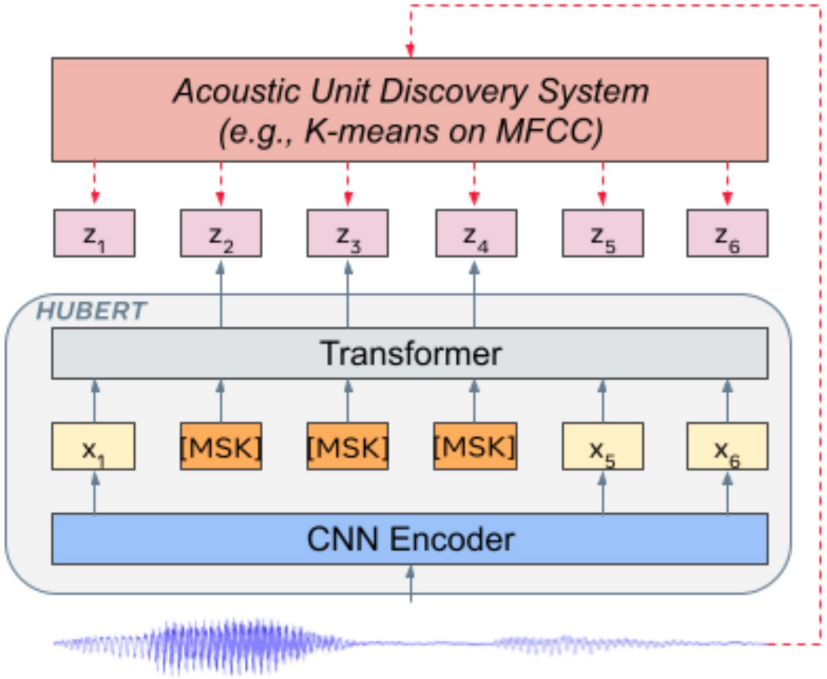


”CPC invents by itself what to predict”

Masked prediction models

Masked models try to predict discretized representations derived from masked regions of input speech.

E.g.: **Wav2vec2.0** (Baevski et al., 2020) or **HuBERT** (Hsu et al., 2021)



Hsu et al. (2021)

SSL for audio-to-audio: general findings

Initial APC and CPC studies: emerging latent representations linearly separable in terms of *phonemic categories* and *speaker IDs*

Current status of audio-based SSL algorithms:

- Extremely well performing features for **phoneme recognition**, **ASR**, **keyword spotting**, **query by example**, **speaker identification**, **speaker diarization**, **intent classification**, **slot filling**, **emotion recognition**…*

	PR	KS	IC	SID	ER	ASR (WER)		QbE	SF		ASV	SD
	PER ↓	Acc ↑	Acc ↑	Acc ↑	Acc ↑	w/o ↓	w/ LM ↓	MTWV ↑	F1 ↑	CER ↓	EER ↓	DER ↓
FBANK	82.01	8.63	9.10	8.5E-4	35.39	23.18	15.21	0.0058	69.64	52.94	9.56	10.05
APC [7]	41.98	91.01	74.69	60.42	59.33	21.28	14.74	0.0310	70.46	50.89	8.56	10.53
modified CPC [34]	42.54	91.88	64.09	39.63	60.96	20.18	13.53	0.0326	71.19	49.91	12.86	10.38
HuBERT Large [35]	**3.53**	95.29	**98.76**	**90.33**	**67.62**	**3.62**	**2.94**	0.0353	**89.81**	**21.76**	5.98	5.75

- **Facebook Wav2vec2.0**: 8.6% WER in ASR with only 10 min of speech with orthography (Zhang et al., 2020).

*SUPERB benchmark: https://superbbenchmark.org/leaderboard

Language learning with audiovisual predictions

Predictive tasks: cross-modal

Infants not only hear speech, but otherwise sense the world around them.

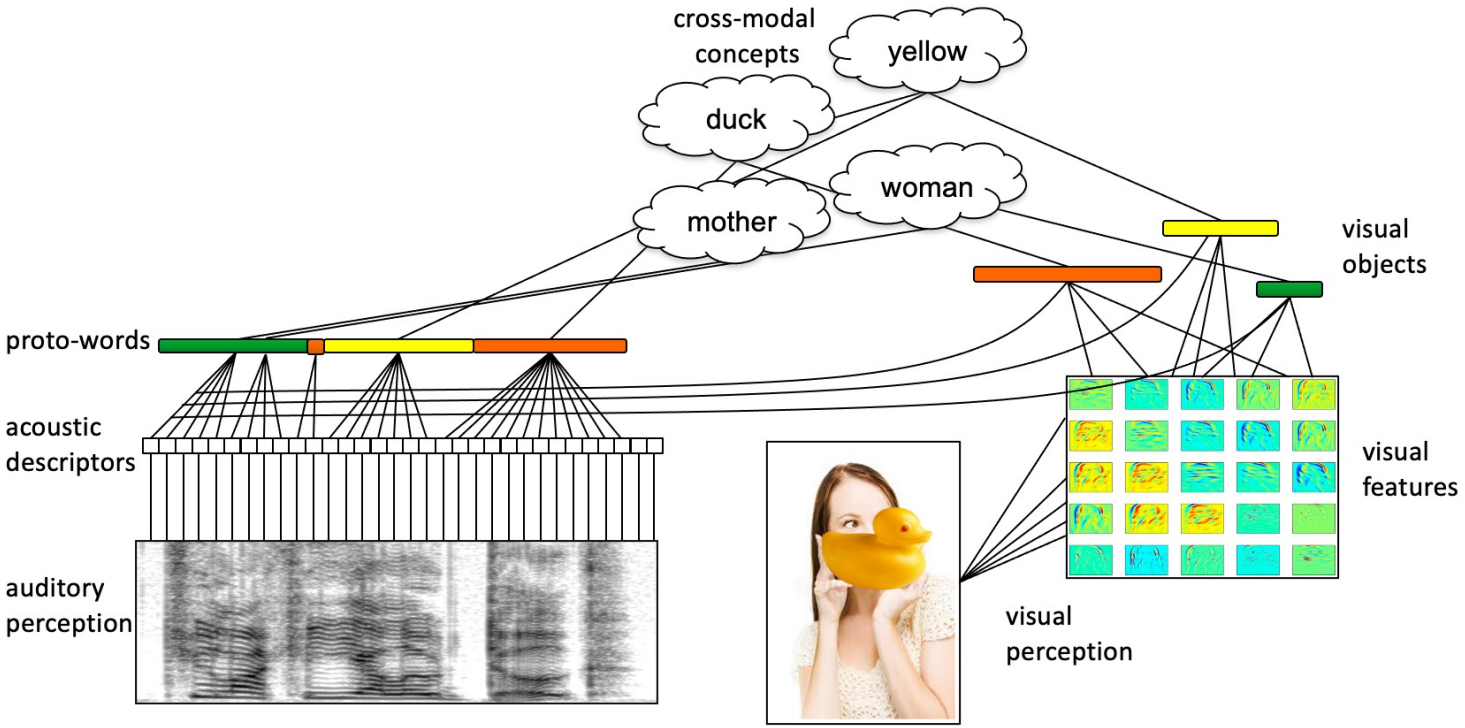
Caregiver speech is not random w.r.t. context, but often relates to concrete objects and events in the situation.

Perhaps multimodal input helps to

1) overcome difficulties in acoustic pattern discovery
2) directly allows connecting words with their referents

Predictive tasks: cross-modal

Räsänen & Rasilo (2015): There's no need for separate word segmentation and word meaning acquisition processes if audiovisual input is available (both in theory and practical simulations).



Predictive tasks: cross-modal

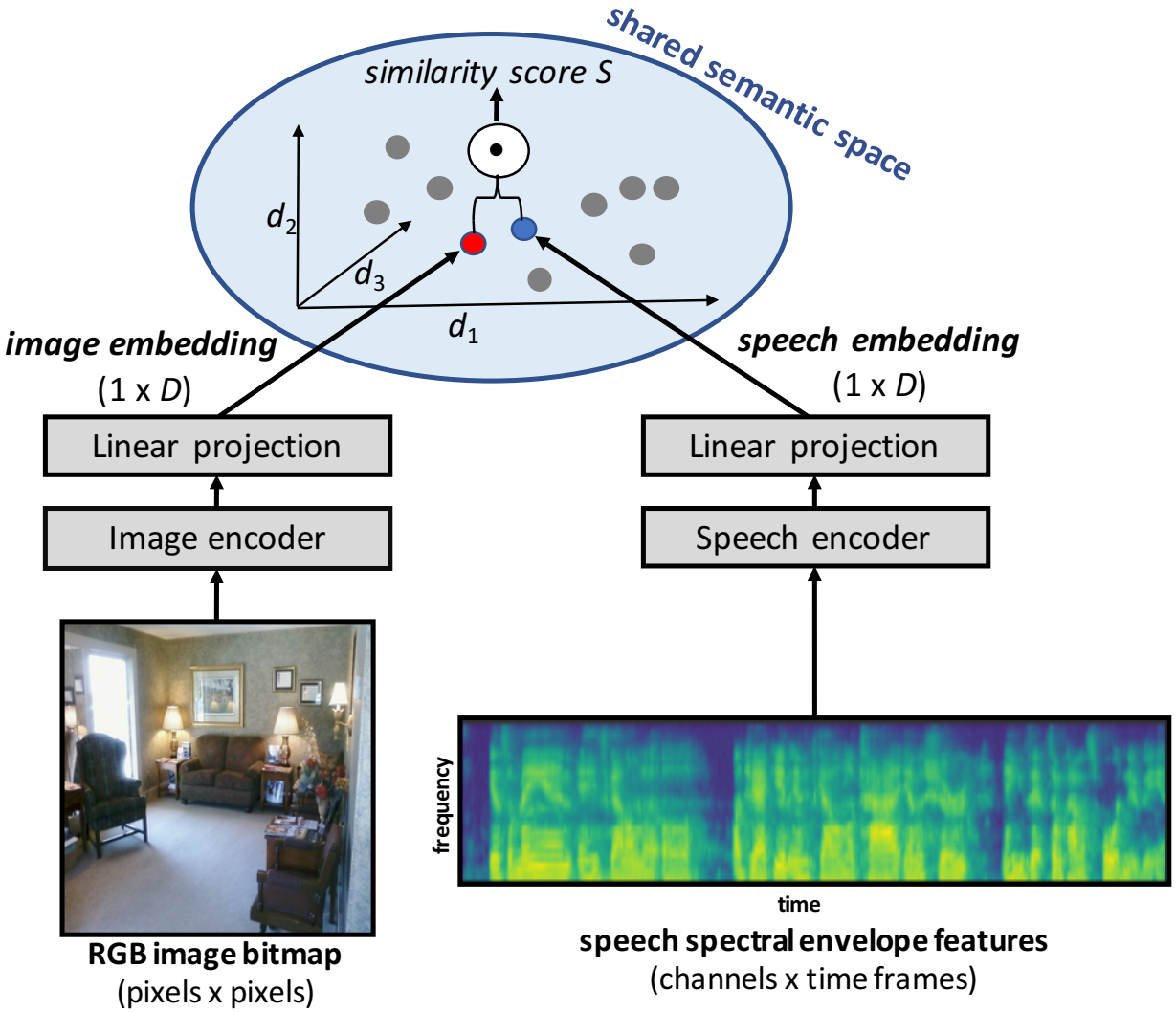
Potential advantages:

- Direct optimization of speech perception w.r.t. its ultimate utility (= how speech predicts the world).
- Visual perception (e.g., object permanence) develops ahead of speech perception → (weak) labels for speech input?

Potential cons:

- Referential ambiguity.
- Real-world data sparse compared to auditory learning.
- Can only explain acquisition of visible/tanglible concepts (cf., e.g., function words).

Visually-grounded speech models (VGS)



VGS models: digging inside

Once trained, we can analyze latent activations in the VGS audio encoders:

- Alishahi et al. (2017): **phonological information** encoded in early hidden layers.
- Chrupala et al. (2017), Merkx et al. (2019): **individual words can be decoded** from hidden layers.
- Havard et al. (2019), Merkx et al. (submitted): **temporal word activation patterns similar to human data**.
- Räsänen & Khorrami (2019): **phonemic and lexical learning** from head-mounted camera data from real infants at home.
- Harwath et al. (2020): **early VQ layers encode phonemic** and **deeper VQ layers lexical units**.
- Khorrami & Räsänen (2021): **phonemic, syllabic and lexical representations emerge concurrently** and largely independently of network architecture.

Joint models with audio-to-audio and audio-to speech predictions?

ZeroSpeech 2021

ZeroSpeech challenge:

- Track 1: auditory learning (Dunbar et al., 2021)
- Track 2: audiovisual learning (Alishahi et al., 2021)
- Learned latents tested for *phonetic*, *lexical*, *syntactic*, and *semantic* tasks.

Basic system architecture: speech SSL + (VGS) + VQ + k-means + BERT LM

				Phonetic (Within)		Phonetic (Across)		Lexical	Syntactic	Semantic	Semantic (Weighted)			
#	Author	Budget	Set	clean	other	clean	other	all	in vocab.		synth.	libri.	synth.	libri.
Some of the best auditory models														
22	Tu Anh et al.	3424	dev	3.26%	4.00%	3.81%	5.91%	70.89%	79.81%	59.49%	2.48	7.72	2.48	7.73
			test	3.03%	3.62%	3.83%	5.63%	71.39%	80.19%	59.29%	8.15	3.12	6.80	3.08
25	Adrian Łancucki et al.	60	dev	2.95%	4.50%	3.54%	7.05%	64.05%	70.86%	52.37%	9.39	13.15	9.39	13.15
			test	2.85%	4.44%	3.67%	7.33%	63.85%	70.23%	51.93%	6.91	0.21	3.09	2.05
11	Chorowski *et al.*	60	dev	2.95%	4.50%	3.60%	6.99%	64.36%	74.04%	52.97%	7.75	4.60	7.75	4.60
			test	2.85%	4.44%	3.69%	7.28%	64.15%	72.47%	52.55%	5.15	0.85	0.74	0.20
The best VGS model														
18	Harwath et al.	468	dev	4.23%	5.58%	4.91%	7.87%	67.60%	75.43%	56.67%	23.07	23.10	23.07	23.10
			test	4.24%	5.22%	5.08%	7.91%	67.56%	75.23%	57.40%	15.10	14.32	17.99	12.78

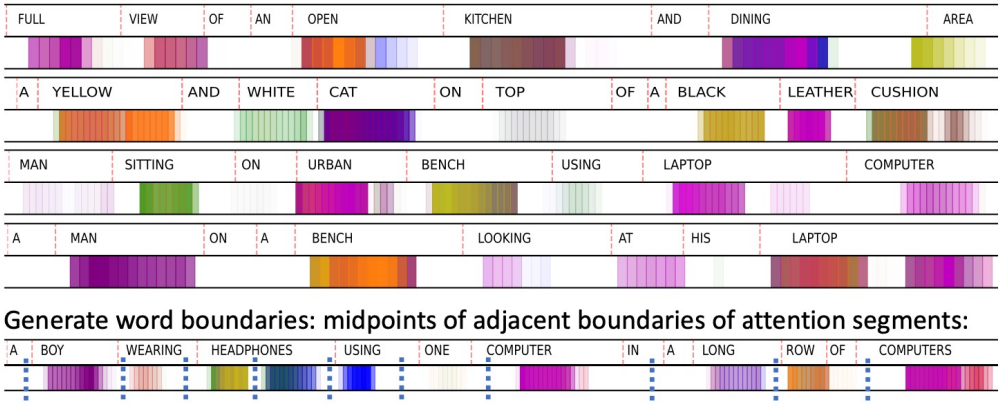
Peng & Harwath (2022)

A model combining unimodal SSL followed by multimodal VGS training.

The key finding:

- Speech ***attention heads discover words and word boundaries***

  …but ***only when trained with visual grounding***.



Model	Boundary				
	Precision	Recall	F_1	OS	R-val
ResDAVEnet-VQ [38]	10.42	50.96	17.30	38.88	-250.77
W2V2 [39]	11.52	24.33	15.63	11.12	-33.34
W2V2_{FT} [39]	11.88	24.79	16.06	10.87	-31.10
HuBERT [40]	12.18	24.97	16.37	10.51	-28.26
$\text{HuBERT}_{\text{FT}}$ [40]	11.90	25.81	16.29	11.68	-36.72
FaST-VGS [28]	28.99	26.17	<u>27.51</u>	-9.72	<u>40.10</u>
FaST-VGS+ [41]	22.66	27.86	24.99	22.93	28.54
VG-W2V2	18.47	19.78	19.10	7.09	28.86
VG-W2V2_4	28.15	22.90	25.26	-18.64	39.67
VG-W2V2_5	28.70	25.45	26.98	-11.32	39.94
VG-HuBERT	18.31	18.90	18.60	3.26	29.60
VG-HuBERT_3	35.90	27.03	**30.84**	-24.72	**44.42**
VG-HuBERT_4	28.39	25.64	26.94	-9.70	39.64

Conclusions

Conclusions, 1/2

Unsupervised (“predictive”) learning from auditory and multimodal sensory streams as a potential mechanism for early language development.

- Compatible with the general idea of *predictive brain* in neuroscience.
- Generalization of the various *statistical learning* mechanisms studied in child language research.

Modern ML provides strong learnability proofs within and across modalities.

- Auditory learning produces useful speech features for a range of speech tasks.
- Audiovisual grounding leads to more human-like semantic representations and leads to lexical segmentation.

Conclusions, 2/2

The relationship between state-of-the-art ML algorithms and human infant learning still unclear.

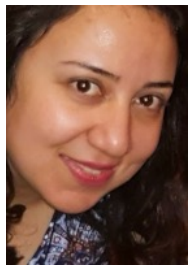
- Mismatch in training data
- Mismatch in evaluation practices
- Both aspects work in progress (and another story)

The end

Speech and Cognition

Speech and Cognition research group

(ps. We're hiring a postdoc!)



Recent funding:

Tampere University

DIGI hum

ACADEMY OF FINLAND

Aalto University

Trans-Atlantic Platform
Social Sciences and Humanities

KONE FOUNDATION