gipsa-lab

Grenoble | images | parole | signal | automatique | laboratoire

# SELF-SUPERVISED LEARNING OF THE RELATIONSHIPS BETWEEN SPEECH SOUNDS, ARTICULATORY GESTURES AND PHONETIC UNITS

*Marc-Antoine Georges, Jean-Luc Schwartz, Thomas Hueber*

Scientific day - AFCP/AFIA - Avignon 2023

cnrs | Grenoble INP | Université Grenoble Alpes | UMR 5216

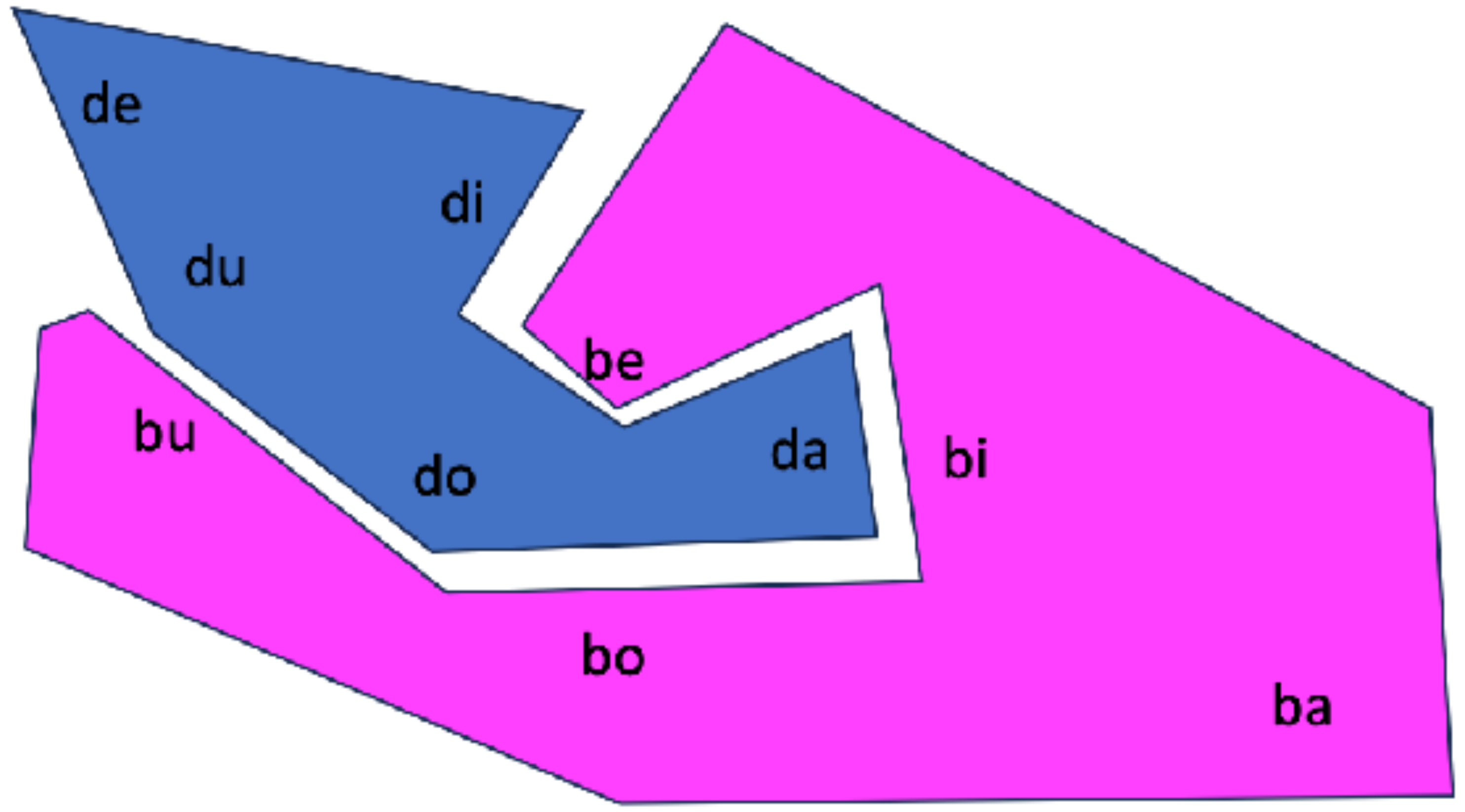
# Introduction

**/a/ - /t/ ?**

**/b/ - /u/ ?**

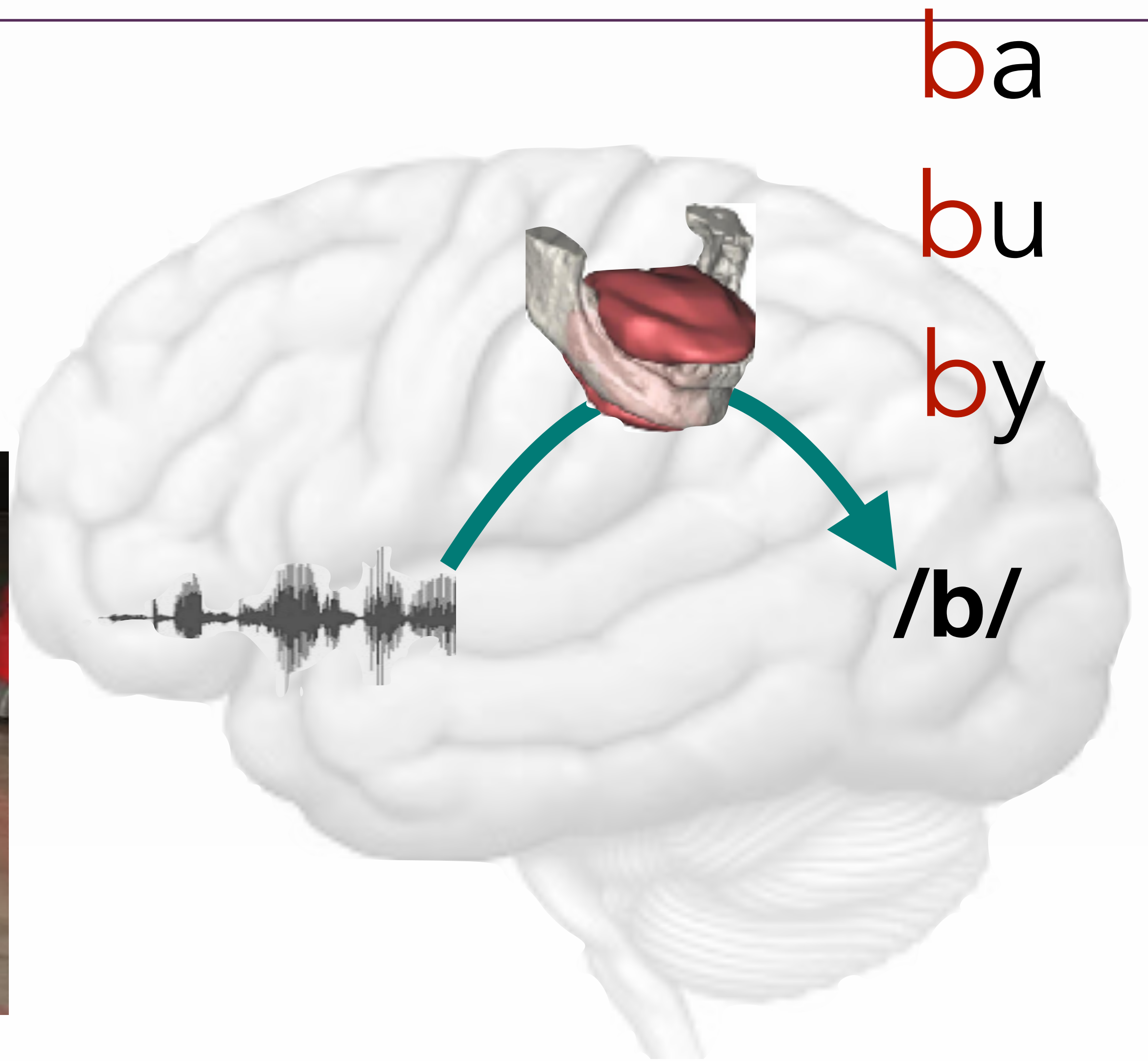
**/i/ - /p/ ?**

- The acquisition of phonology
  - Discovering discrete and invariant units from noisy acoustic inputs
    - inter speaker variability
    - intraspeaker variability (**coarticulation,** Liberman (e.g., 1957))

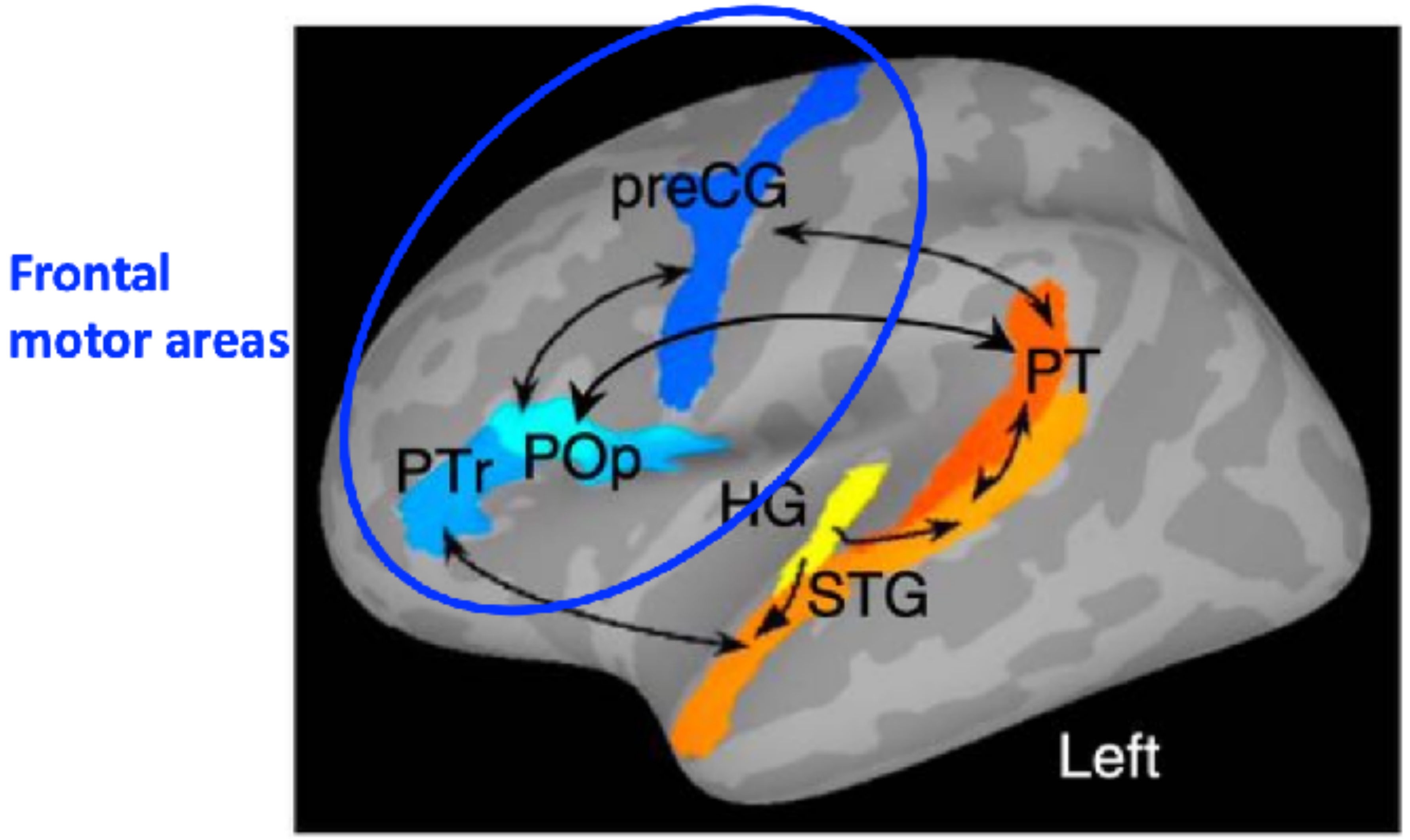



Source: Anne Vilain (GIPSA-lab)

- Invariance can be found in the articulatory domain !
- Motor & perceptuo-motor theories of speech perception (Liberman and Mattingly, 85) (Schwartz et al., 2012)
  - Internal motor simulation: transforming an auditory input into a set of motor commands
  - Efficient when learning a new sound + in adverse conditions
- Neuro-physiological correlates (Pulvermuller et al., 2006), (Sato, Tremblay, & Gracco, 2009), (D'Ausilio et al. 2011; Skipper et al. 2017; Möttönen et al. 2013; Murakami et al. 2015; Du et al. 2016, etc.)

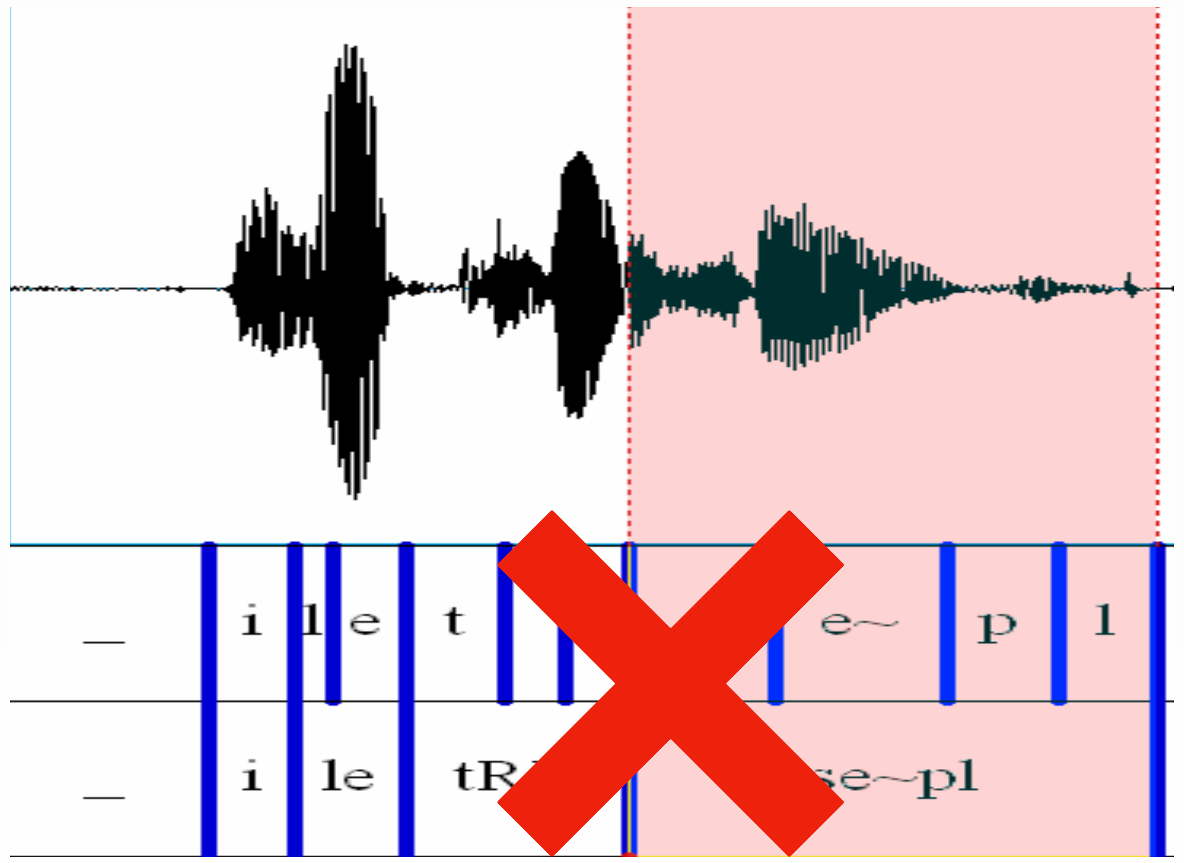



Ex: Du et al. 2016

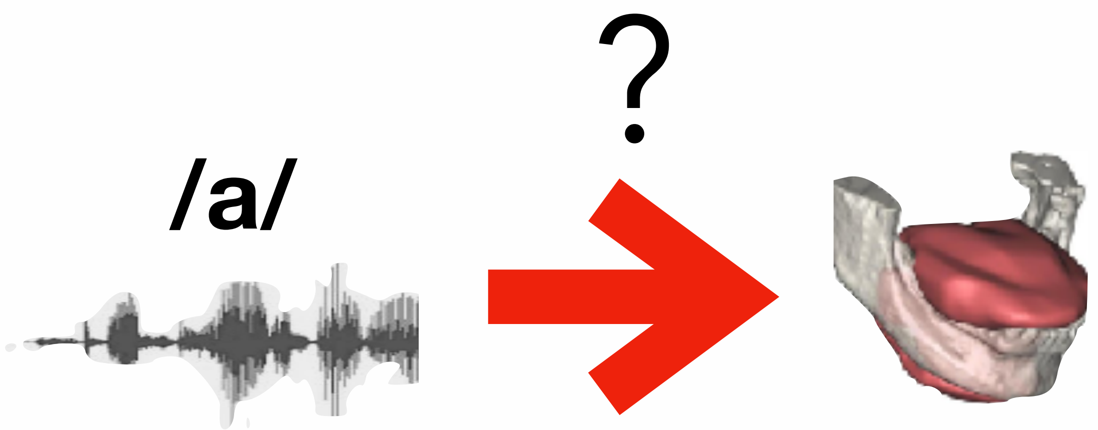


# A weakly-supervised learning

- Children seem to learn the « sound-gesture-speech units » relationships in a **weakly supervised manner**
  - no labeling of the acoustic input
  - no access to the target configuration of the vocal tract for a given input sound (children learn the acoustic-to-articulatory mapping)




  - Acoustic-articulatory mapping, a ill-posed problem
    - non-linear & many-to-one (Atal, 1978), (Qui & Carreira-Perpiñán, 2007), (Neiberg et al, 2008)



# Computational model of speech learning

- Studying speech learning using computer-based simulations

- Explicit integration of speech production knowledge (Moulin-Frier et al., 2014), (Rasilo et Räsänen, 2017) (Philippsen et al., 2021), (Pitti et al., 2021)

  - But most of them are built from (and tested on) simple linguistic material, sometimes synthetic

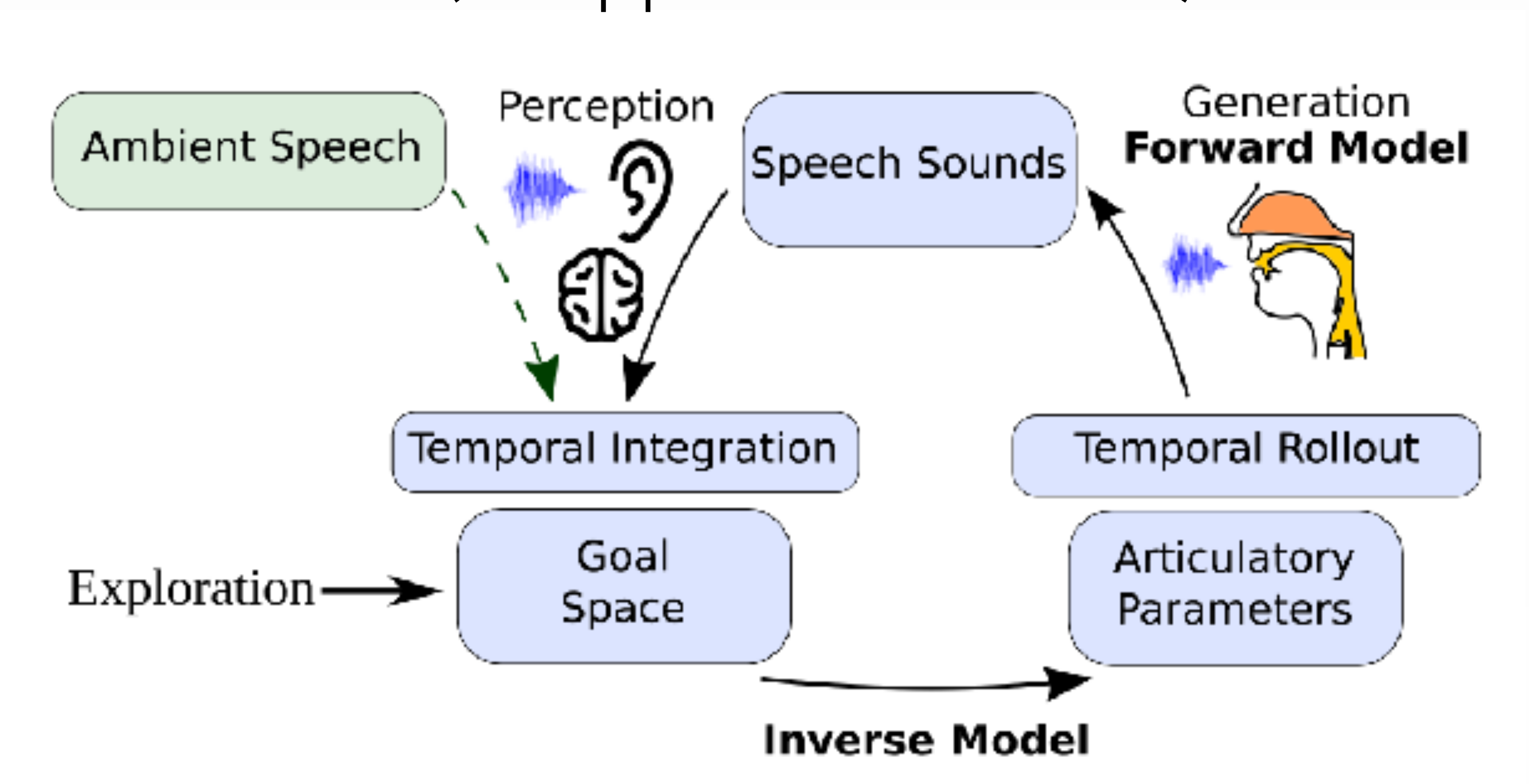
- Deep learning approach exploiting massive data

  - (Dupoux, 2016) « *constructing scalable computational systems that can, when fed with realistic input data, mimic language acquisition as it is observed in infants* ».

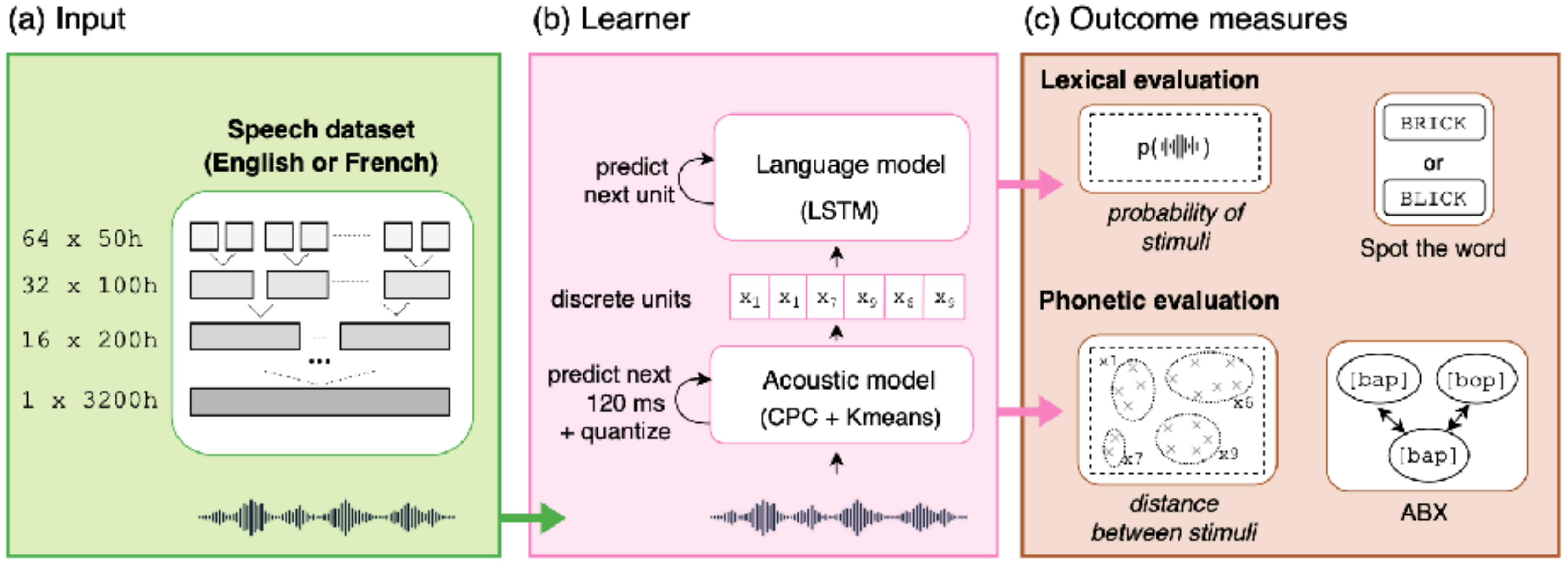
  - STELLA model (Lavechin et al., 2023) able to learn phonological units, but only from clean audio data

  - **No information about speech production**

(Philippsen et al. 2021)




STELLA Model (Lavechin et al. 2023)



# Our research goal

- Build a computational model of speech acquisition based on deep learning, with explicit knowledge of speech production

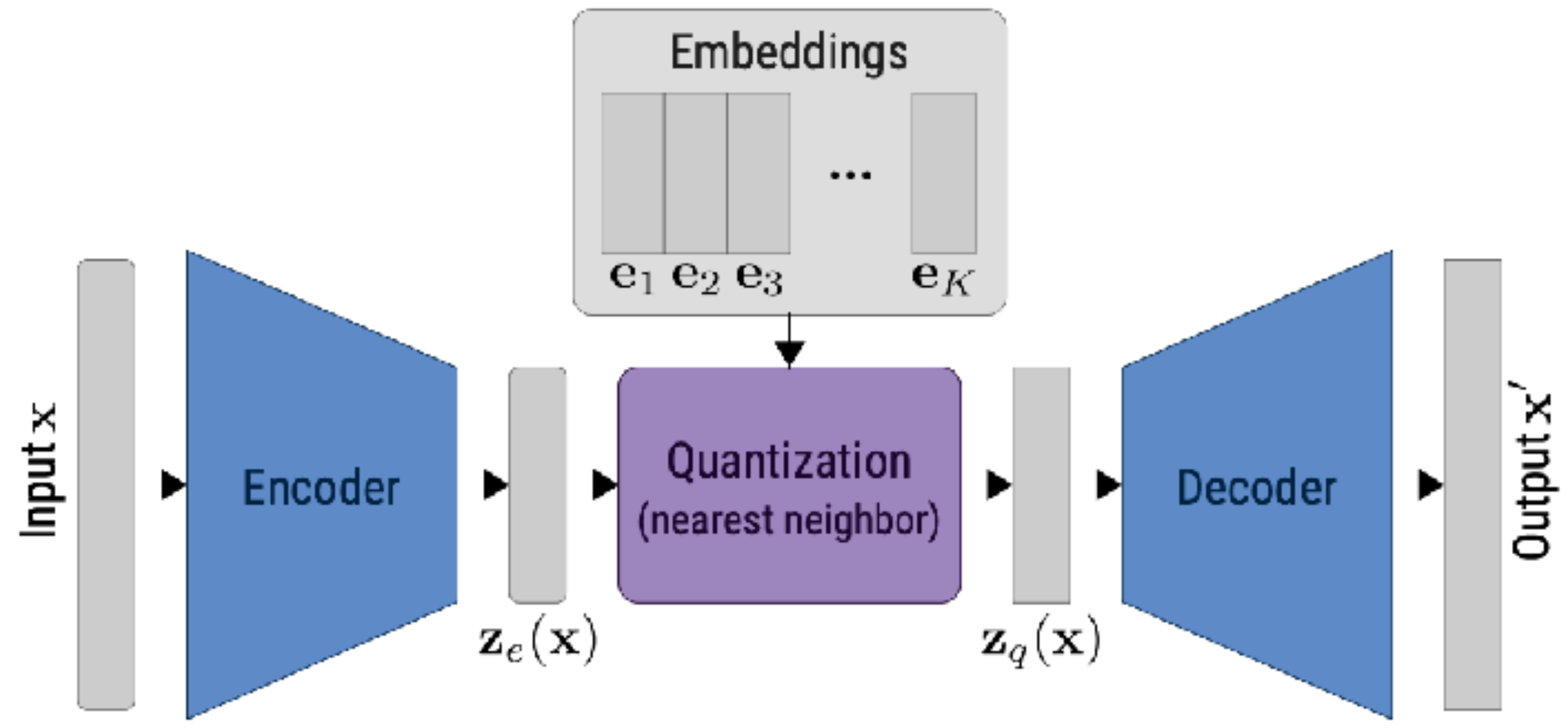
- Research questions
  - How the visual information (aka lip movements) change the embedding of SSL models based on predictive coding? (Hueber et al., Neural Computation, 2020)
  - Does an explicit access to articulatory knowledge improve speech decoding in adverse conditions? (Georges et al., Interspeech 2021)
  - Can prior articulatory knowledge make the learning of phonological units easier? (Georges et al., Interspeech 2022)
  - How the speech-gesture-unit relationship can be learned in a self-supervised manner? (Georges et al., ICASSP 2022)

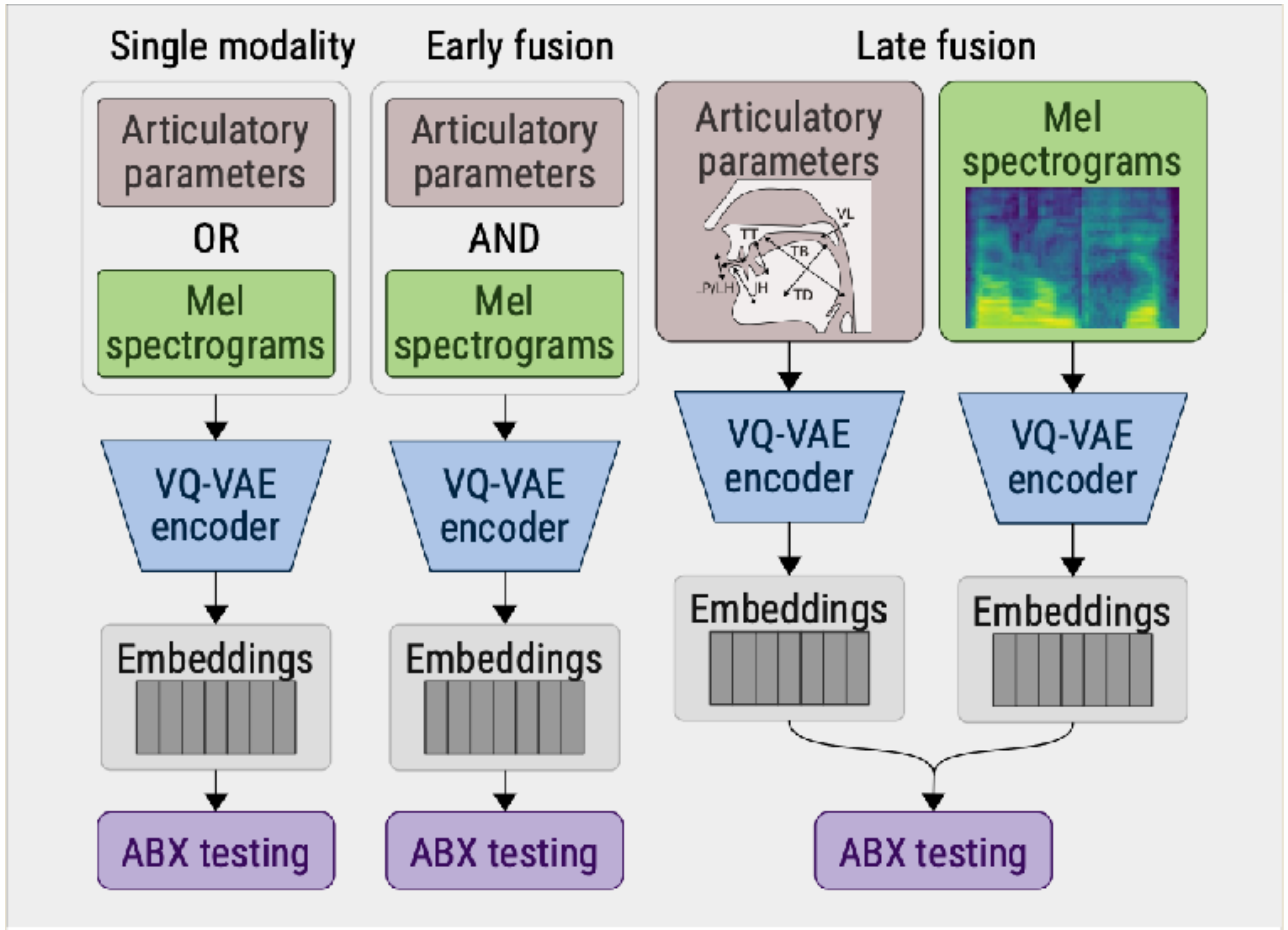
Marc-Antoine Georges PhD

# Role of articulatory knowledge in phonological unit discovery

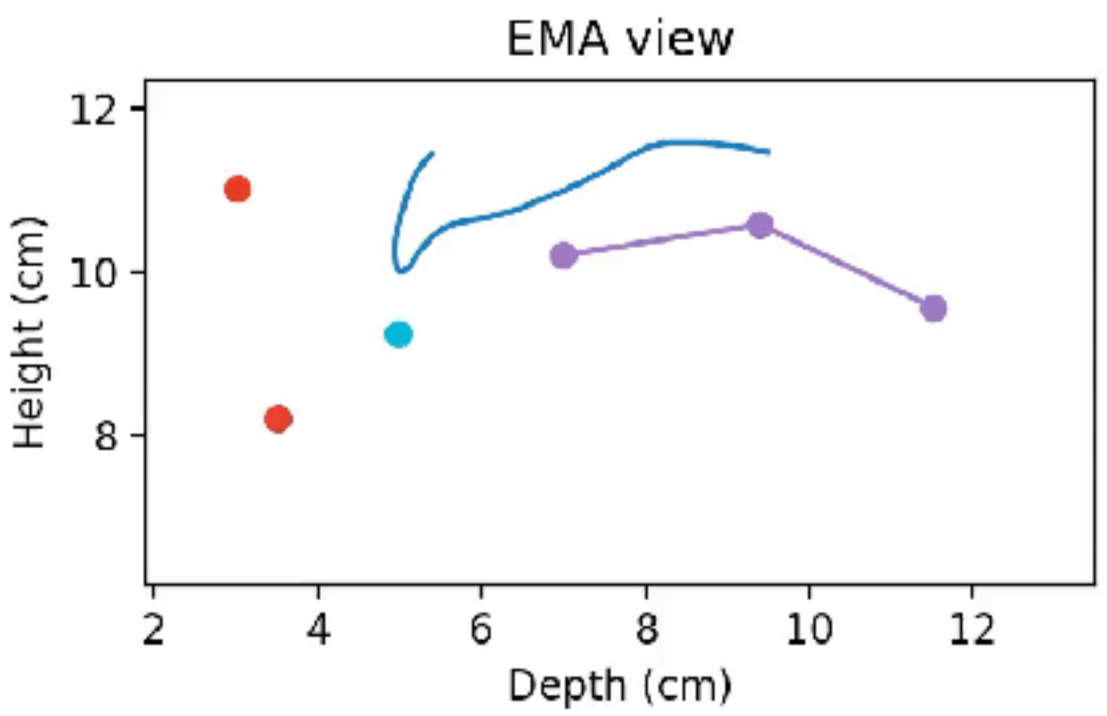
- VQ-VAE ~ VAE but with a **discrete** embedding space
  - Common model used in the Zero-Ressource challenge for unsupervised speech unit discovery (Tjandra et al., 2019), (Niekerk et al., 2020)
- Approach:
  - VQ-VAEs trained either from acoustic / articulatory / acoustic+articulatory data
  - Assessing the phonetic discriminability of the learned embeddings using ABX tests (Schatz et al., 2013)

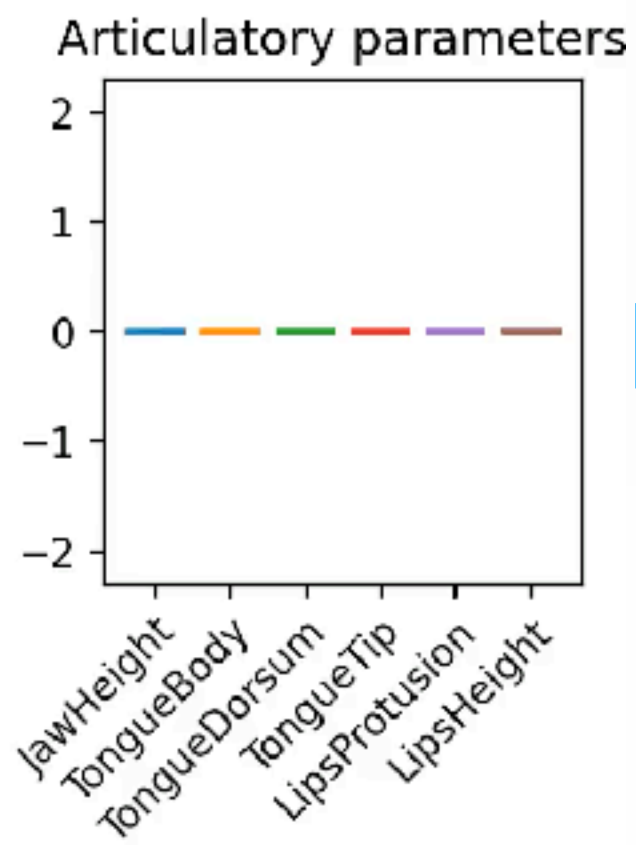
VQ-VAE (van den Oord et al., 2017)

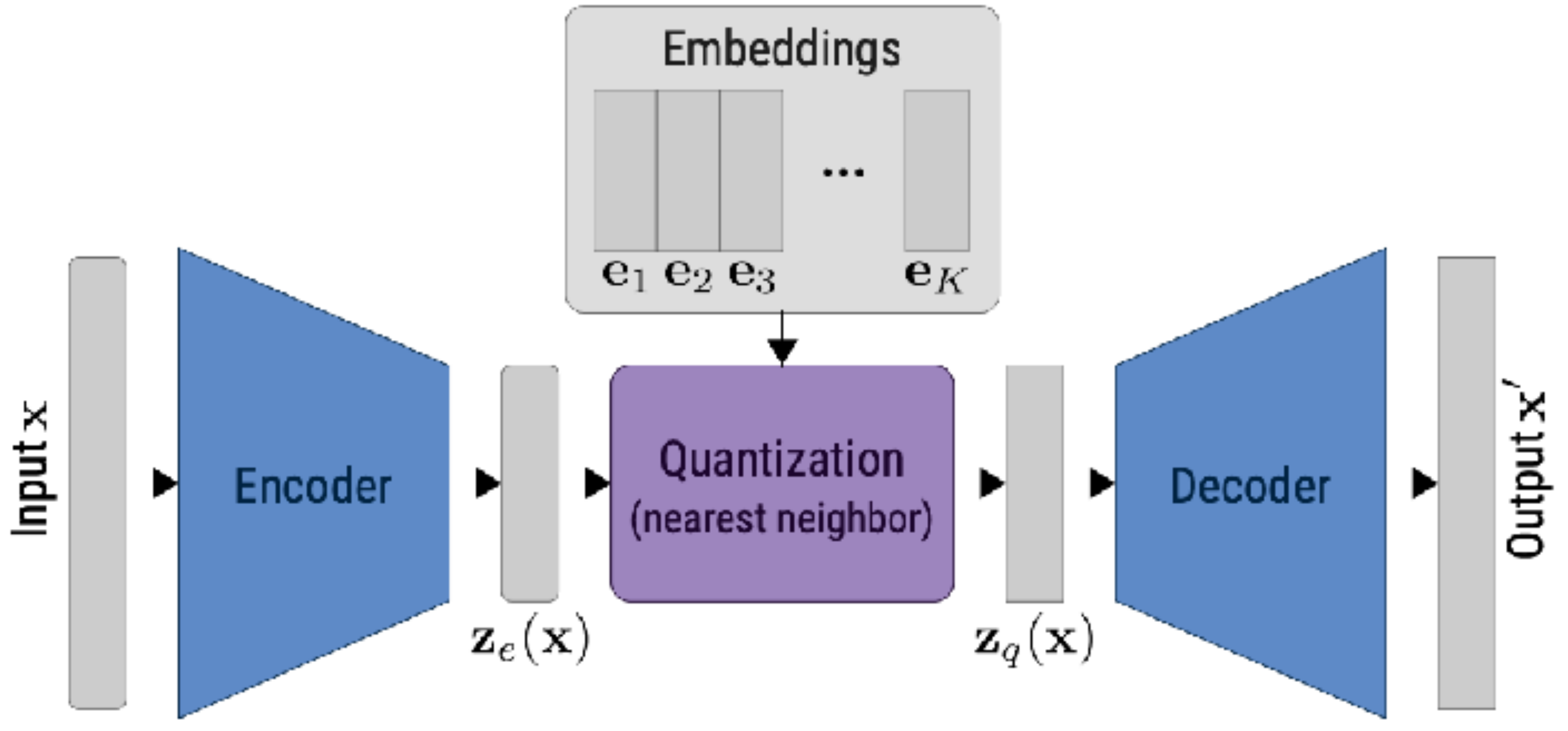


# Role of articulatory knowledge in phonological unit discovery




Datasets: PB2007 & BY2014 (2 French speakers) and MOCHA-TIMIT (7 English speakers), Codebook size K=32,64,128,256, 512

## ABX methodology

Two representations of the same unit should be closer to each other than to any other unit representation

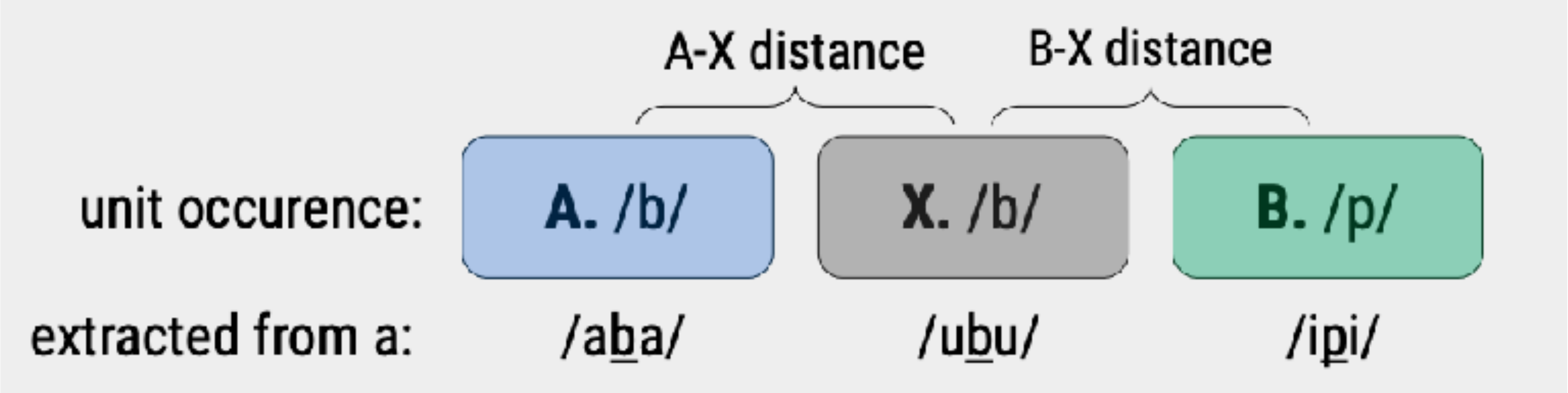
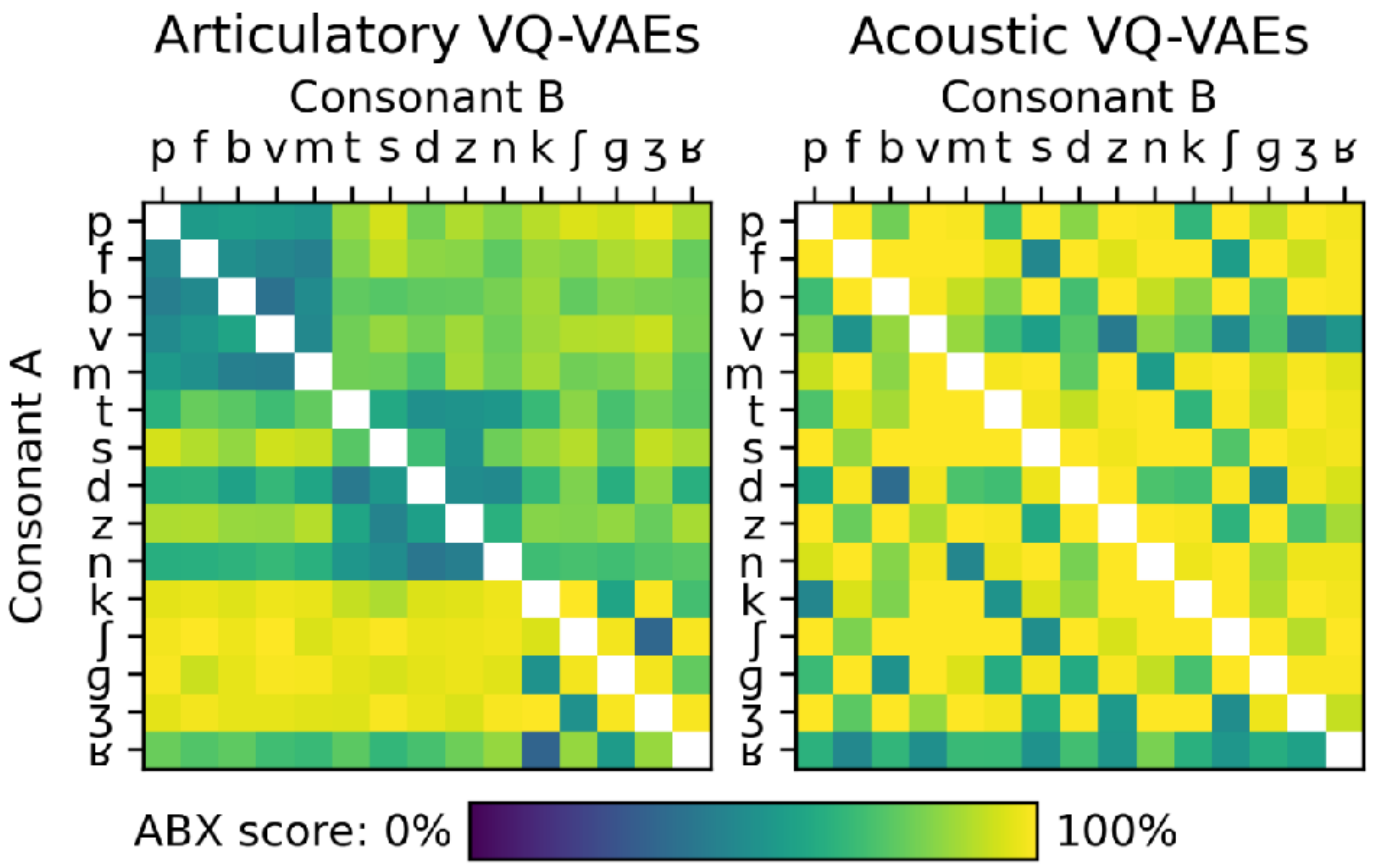



Illustration of an ABX test

# Role of articulatory knowledge in phonological unit discovery



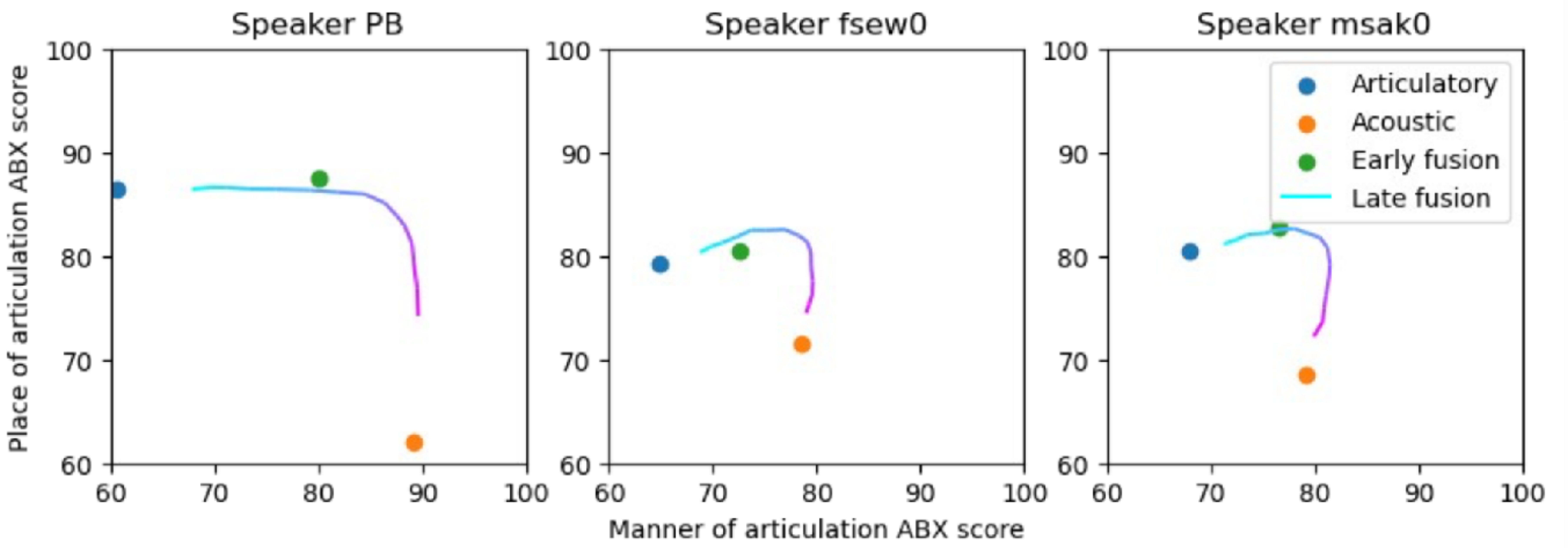

**Structure of the latent space**

Articulatory modality = place of articulation

Acoustic modality = manner of articulation.

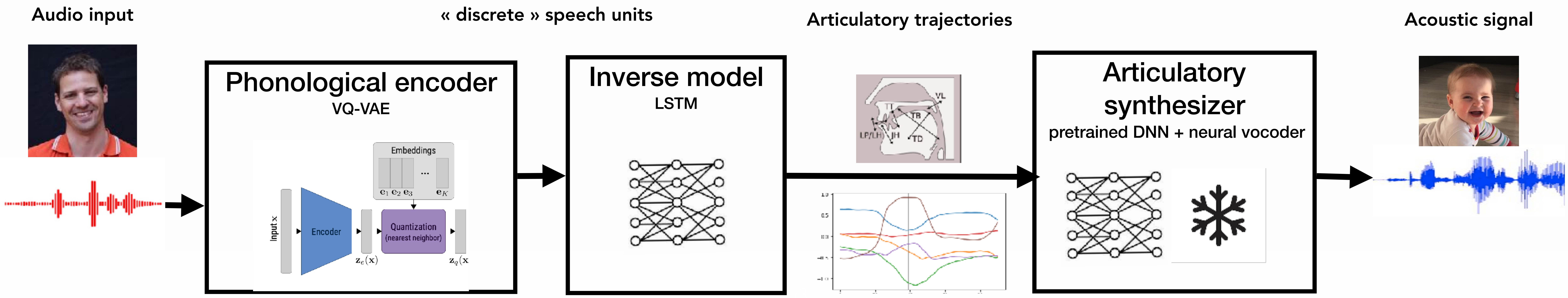
# Role of articulatory knowledge in phonological unit discovery

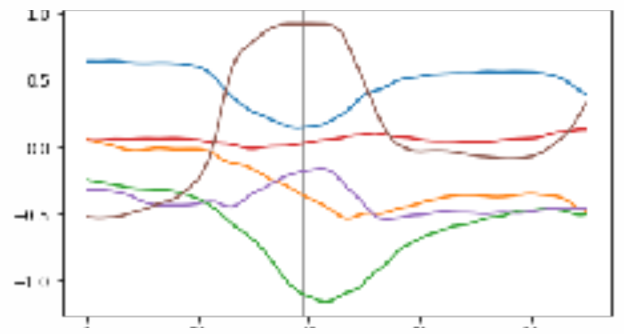
ABX score for the consonants - Place vs. manner of articulation - 3 speakers (out of 9)




Fusion of articulatory and acoustic modalities ➡ better phonetic discriminability

# Computational model of speech learning



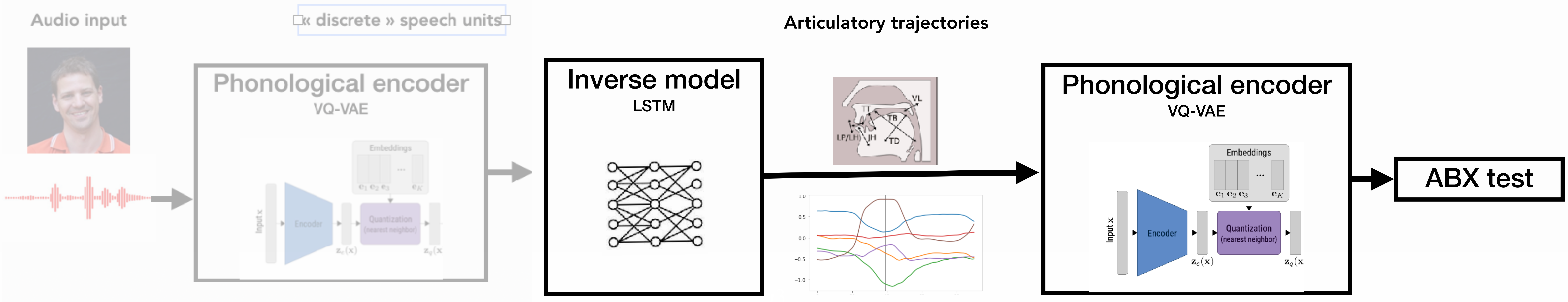

$L = L_{VQVAE} + L_{acoustic}($  ,  $) + L_{jerk}\left(\frac{d^3}{dt^3}\right.$  $\left.\right)$

Synthetic speech is intelligible

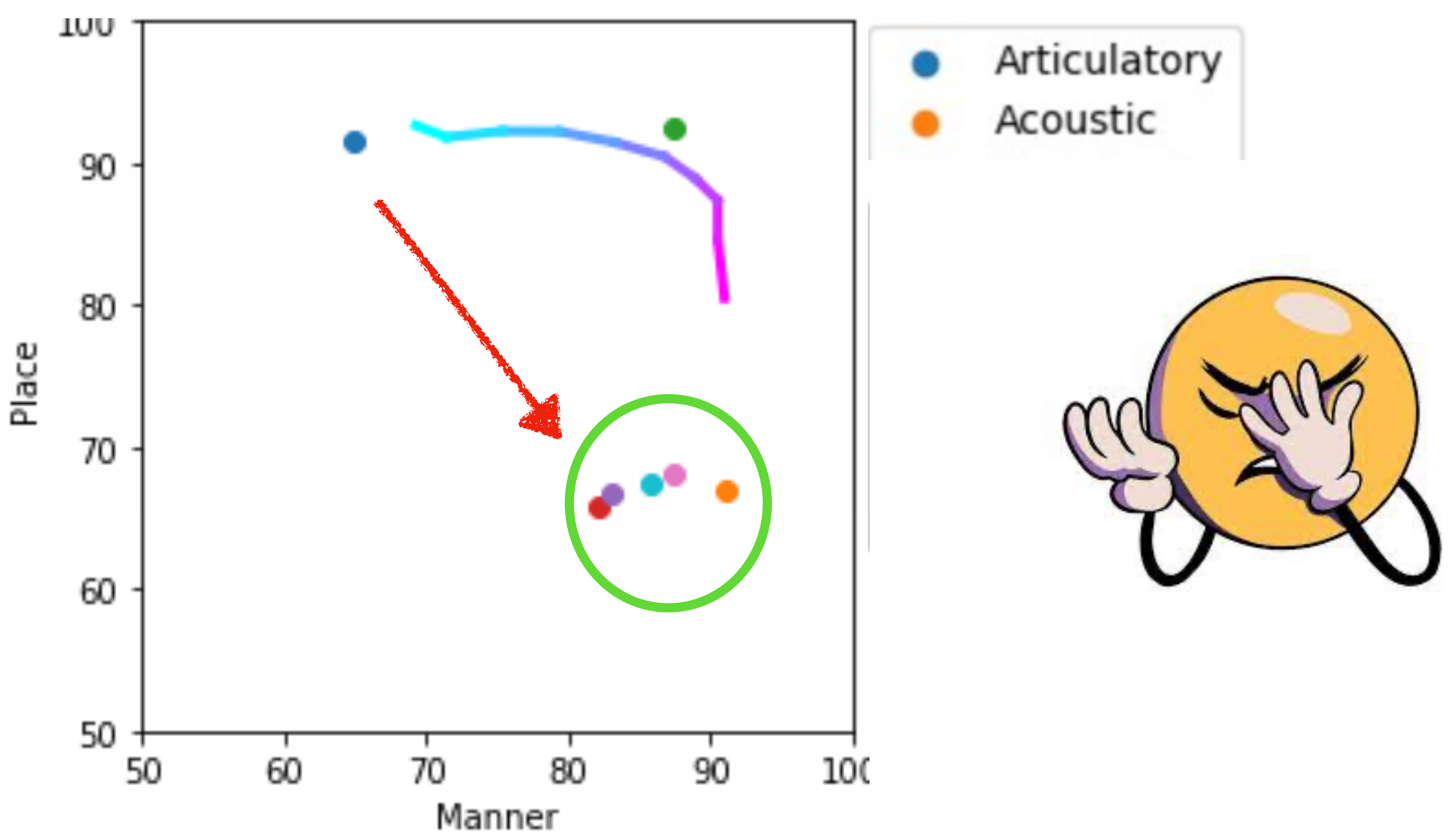
(but only for speakers relatively

close to the reference speaker)

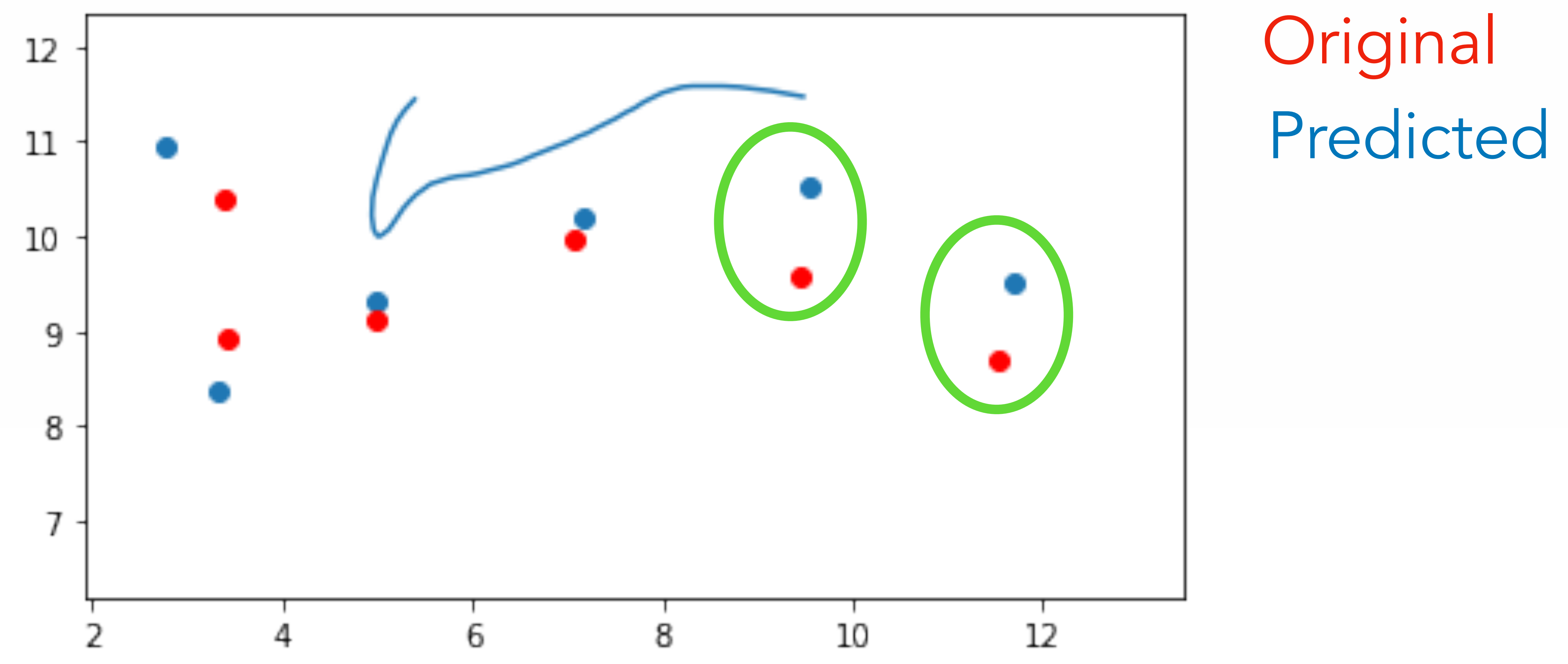
# Computational model of speech learning




## ABX score




## a[b]a - Predicted vs. ground truth vocal tract config.



# Conclusions and perspectives

- Goal : Investigate how a child learn the relationships between phonological unit - speech sound and articulatory gestures
- Approach : Computer-based simulation using deep networks + SSL trained from raw speech datasets
- Main results : Complementary role of articulatory and acoustic knowledge
  - in adverse condition (AR-VAE)
  - for discovering phonological unit (VQ-VAE)
  - current model unable systematically infer plausible articulatory trajectories :-(
- Perspectives
  - Introducing biomechanical constraints in the inverse model (PhD Angelo Ortiz, co-dir E. Dupoux)
  - Investigating the role of babbling strategies (Post-doc Marvin Lavechin)
  - Introduce a weak supervision signal (RL, multimodal input)

# The end


- Georges M-A, Schwartz J-L, Hueber, T., "Self-supervised speech unit discovery from articulatory and acoustic features using VQ-VAE", Proc. of Interspeech, 2022,
- Georges M-A, Diard, J., Girin, L., Schwartz J-L, Hueber, T., "Repeat after me: self-supervised learning of acoustic-to-articulatory mapping by vocal imitation", Proc. of ICASSP, pp. 8252-8256, 2022
- Georges M-A, Girin L., Schwartz J-L, Hueber, T., "Learning robust speech representation with an articulatory-regularized variational autoencoder", Proc. of Interspeech, pp. 3345-3349, 2021
- Hueber, T., Tatulli, E., Girin, L., Schwartz, J-L., "Evaluating the potential gain of auditory and audiovisual speech predictive coding using deep learning", Neural Computation, vol. 32 (3), pp. 596-625.