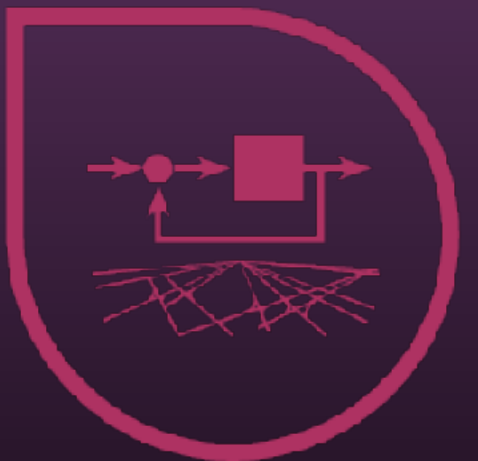




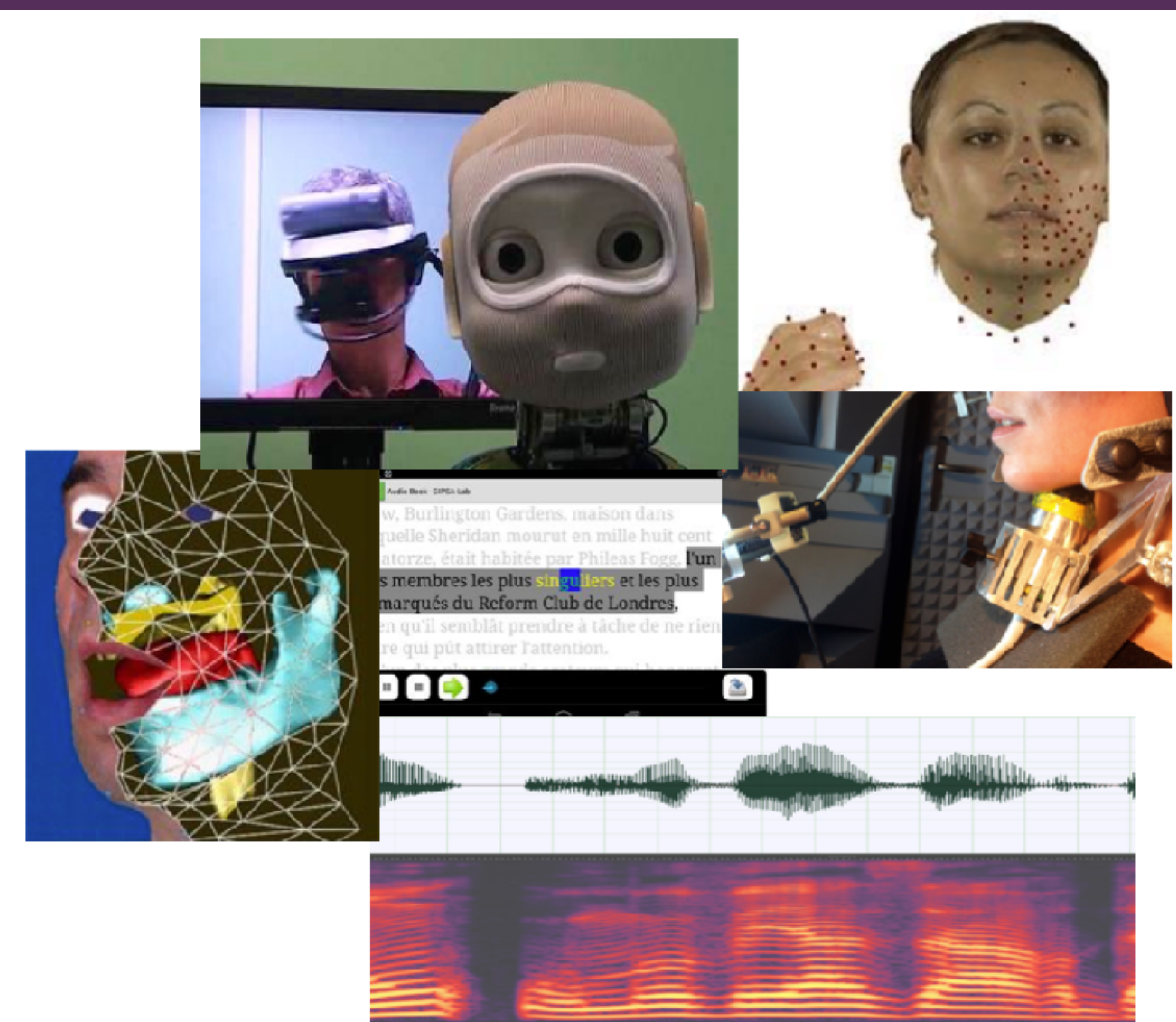
MODELING SPEECH ACQUISITION USING SELF-SUPERVISED MACHINE LEARNING, A FOCUS ON THE ACOUSTIC-TO-ARTICULATORY MAPPING

Thomas Hueber, GIPSA-lab

Seminar LIS/Talep, 2022

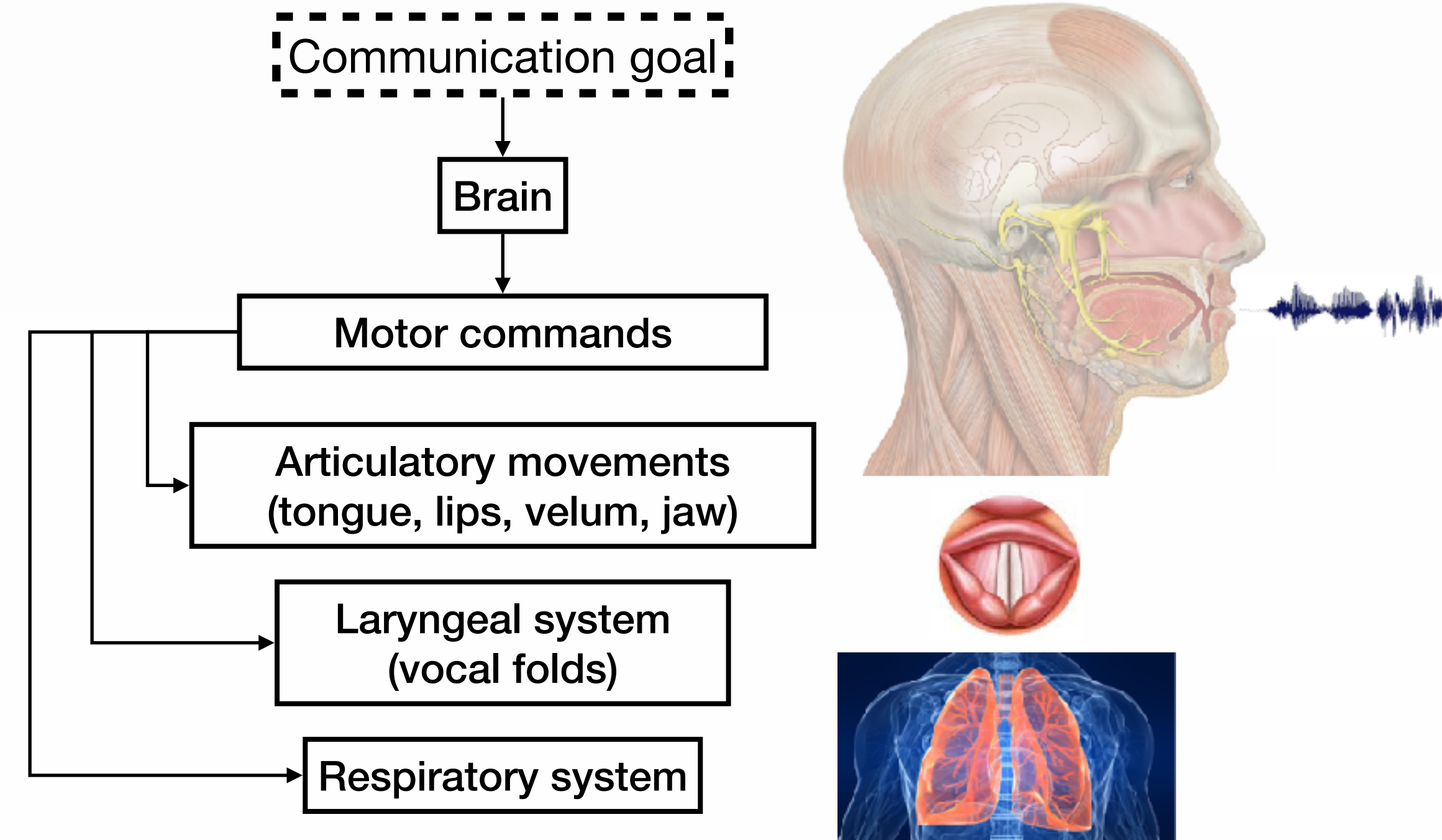
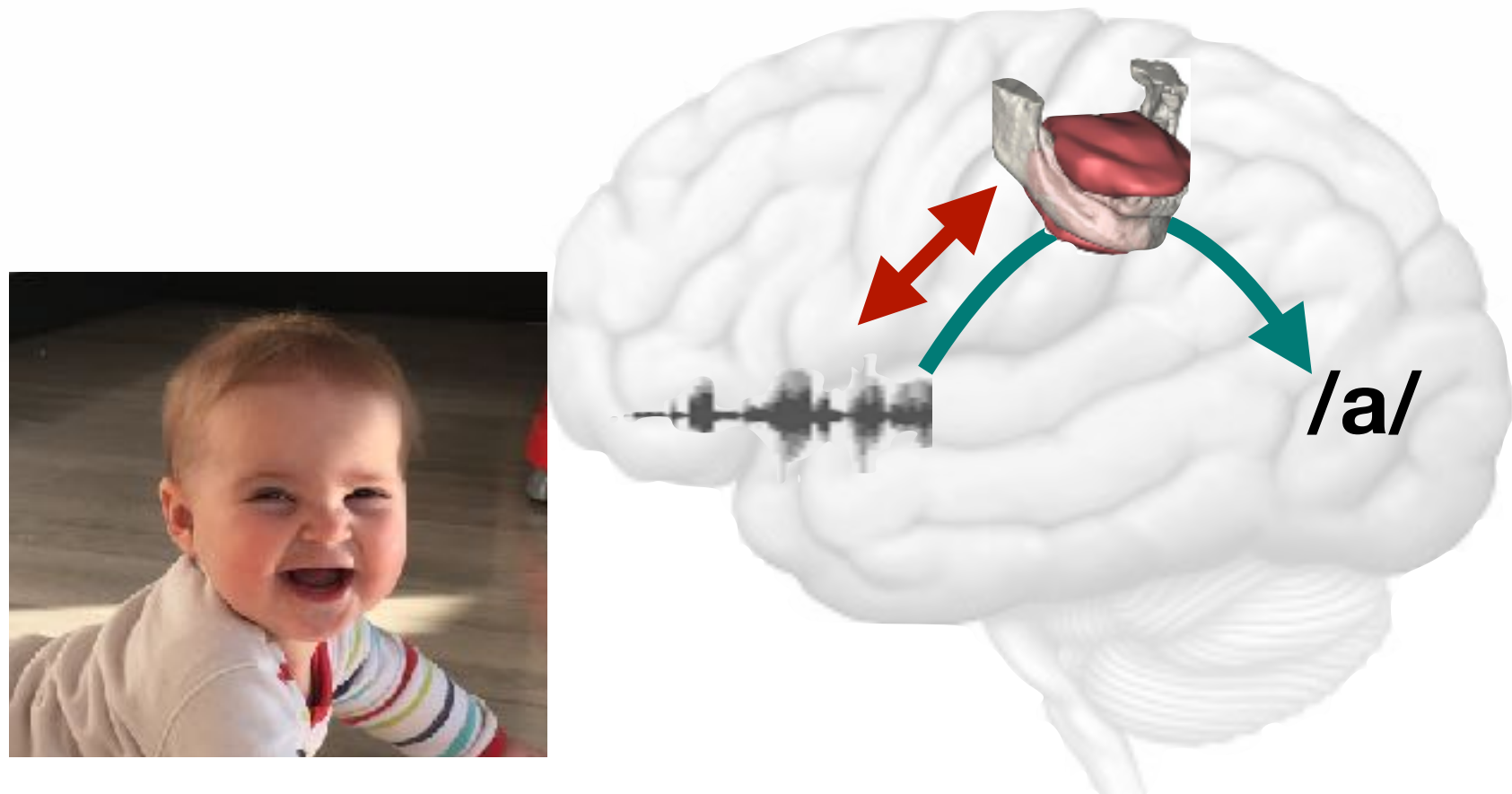


- Automatic speech processing
 - Speech synthesis (prosody modeling, expressivity, low-latency, gesture-based control)
 - Multimodal speech processing (visual speech recognition, Cued-speech)
 - Speech enhancement & source separation
 - **Speech production**
 - Acquisition and modeling of co-verbal signal
 - Conversational agent / Humanoid robot
- 1 PR, 3 DR CNRS (1 emerite), 2 CR CNRS, 2 IR CNRS & 7 PhD candidates
- Involved in the chair « Bayesian Cognition and Machine Learning for Speech communication » of the Grenoble 3IA institute MIAI (GIPSA-lab, LPNC)



Introduction

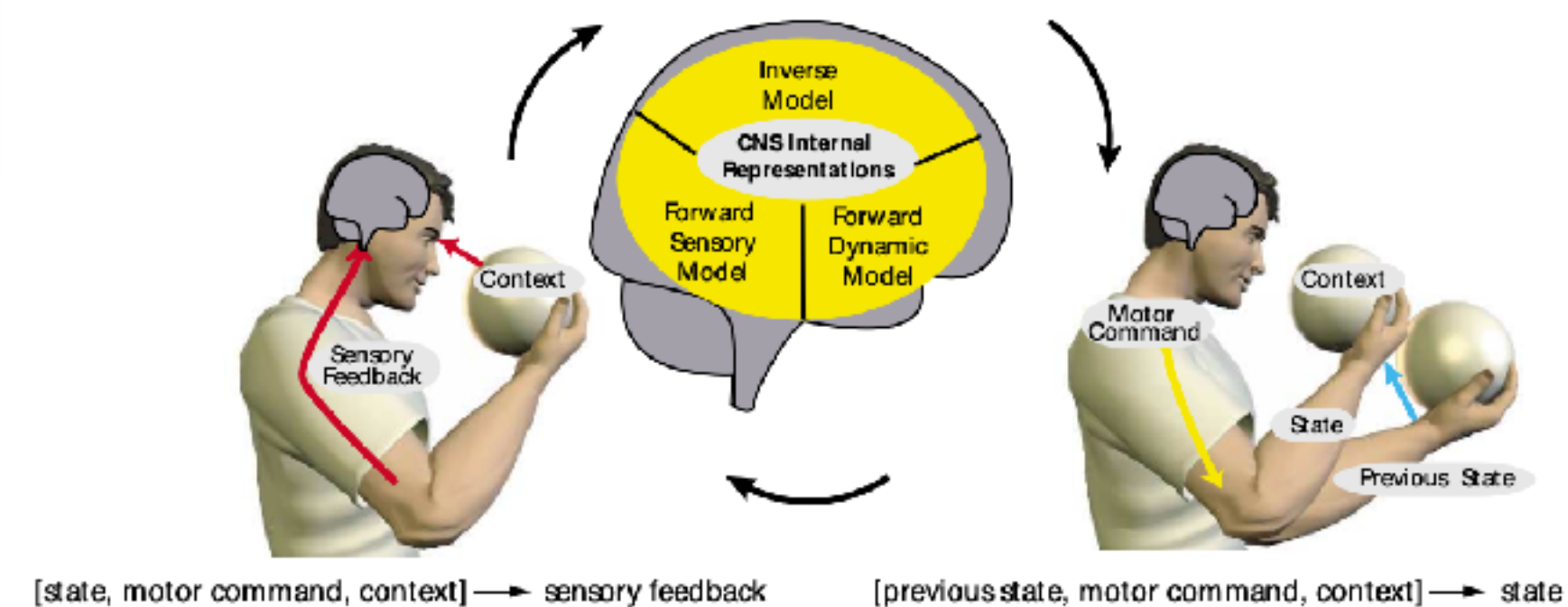
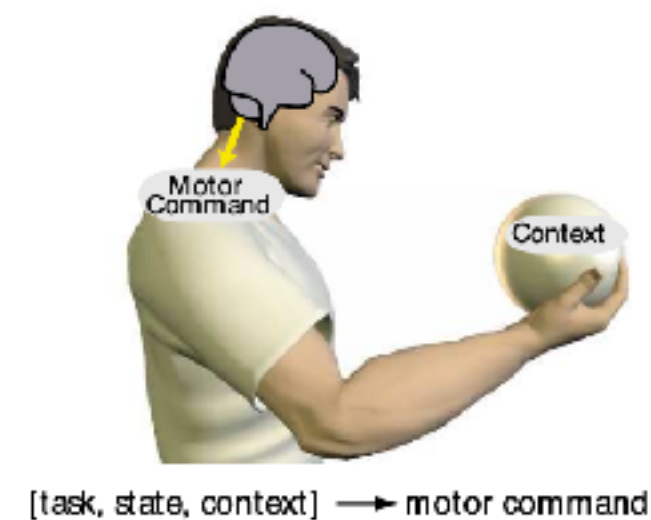
- Speech production is a complex motor process
- Speech and language acquisition = discovering & learning the complex relationships between acoustic/articulatory (motor)/linguistic levels
- Focus on the articulatory level & acoustic-articulatory mapping



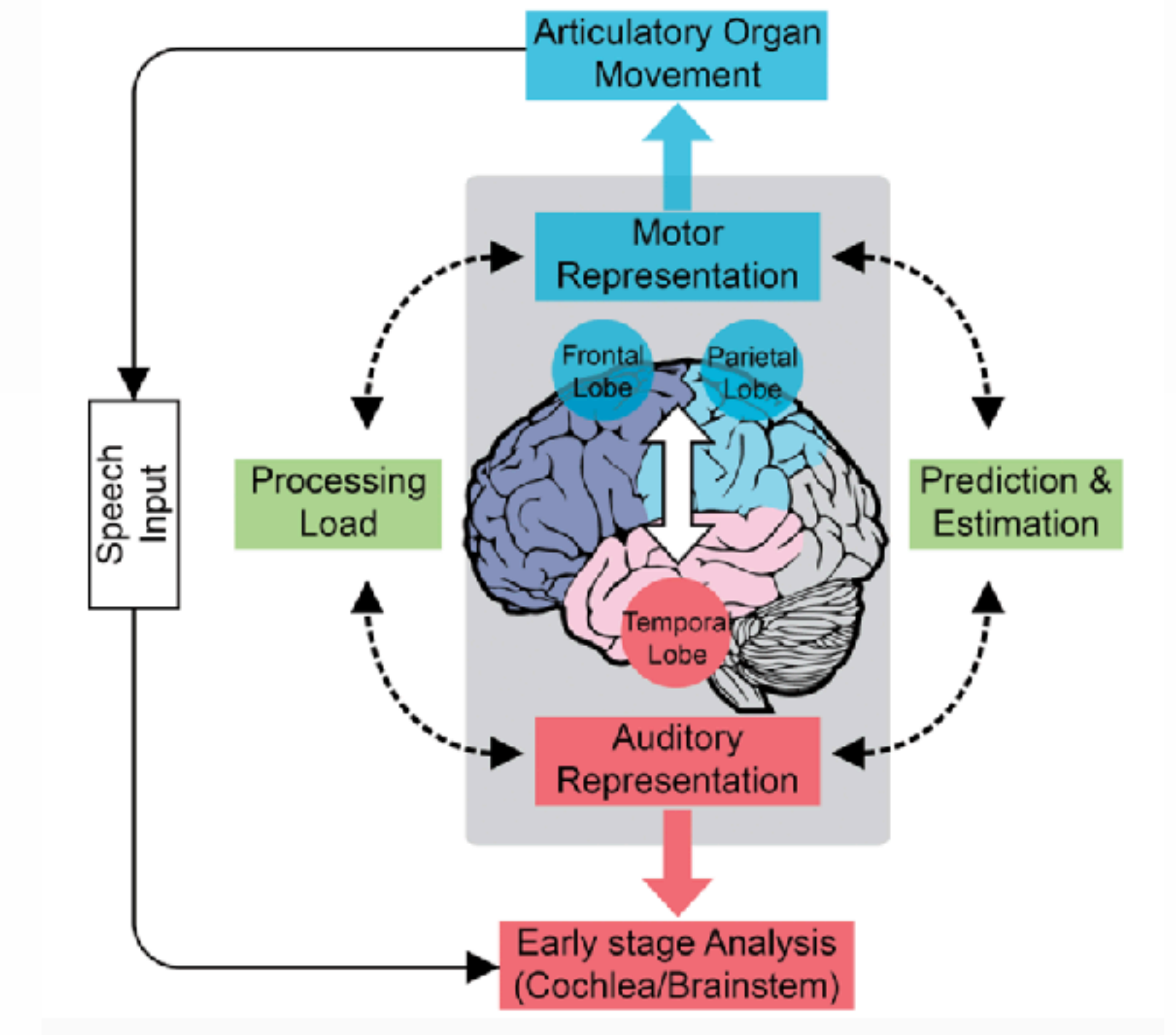
Acoustic-articulatory mapping, a major role in speech perception and production

- Speech perception

- Articulatory input helps decoding speech in adverse conditions
 - Lips (Sumbly and Pollack, 1954; Benoît et al., 1994; Grant and Seitz, 2000), but also tongue (Badin et al., 2008), or tactile information (Treille et al. 2017)
- Motor/Perceptuo-motor theories Liberman and Mattingly, 85) (Schwartz et al., 2012)
 - Speech perception —> transforming of auditory input into a set of motor commands (see Pulvermuller et al., 2006) or (Sato, Tremblay, & Gracco, 2009) for neuro-physiological correlates)



from Wolpert & Ghahramani, Nature, 2000



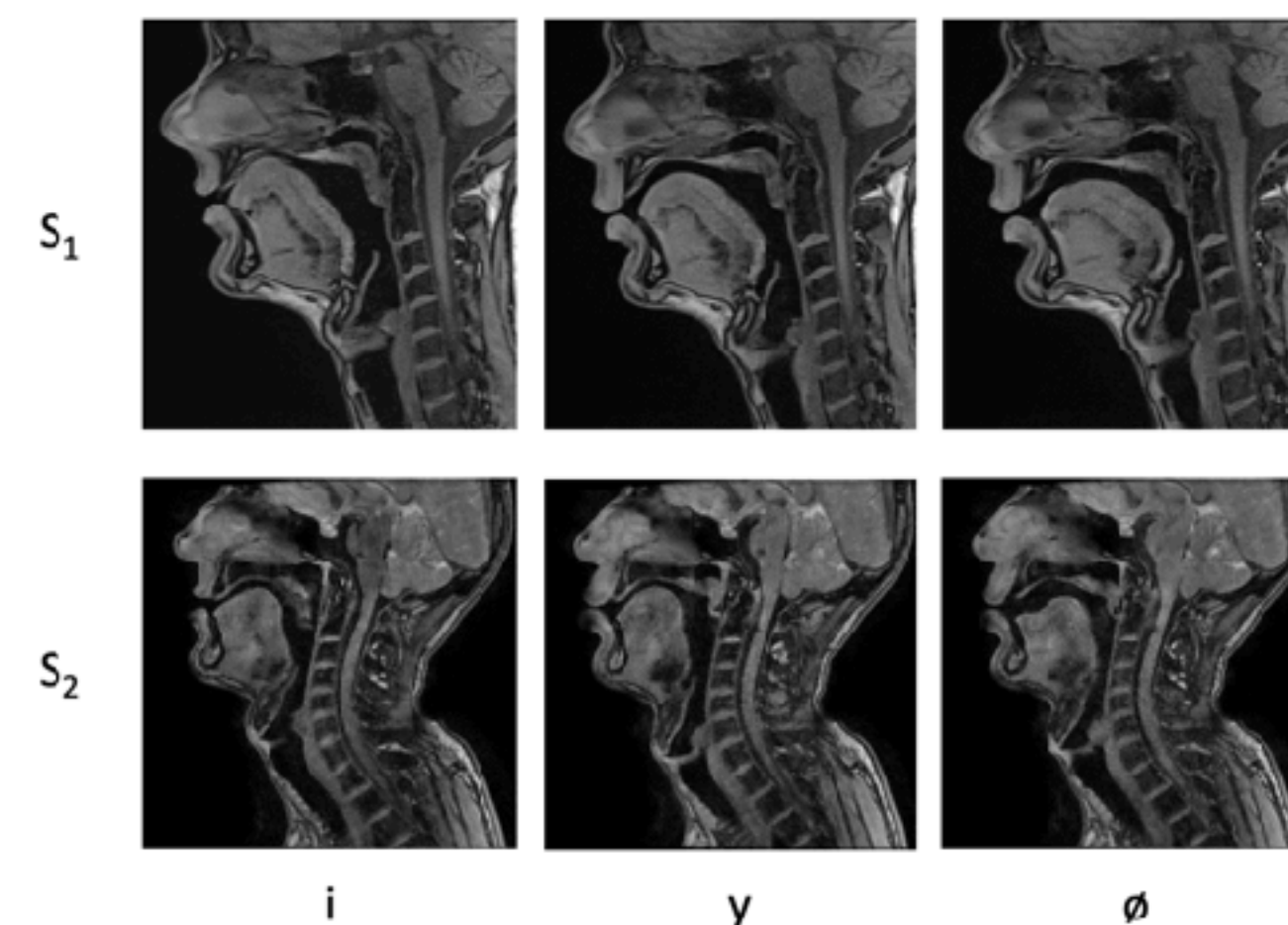
from Wu et al., J. Neuroscience, 2014

- Speech production

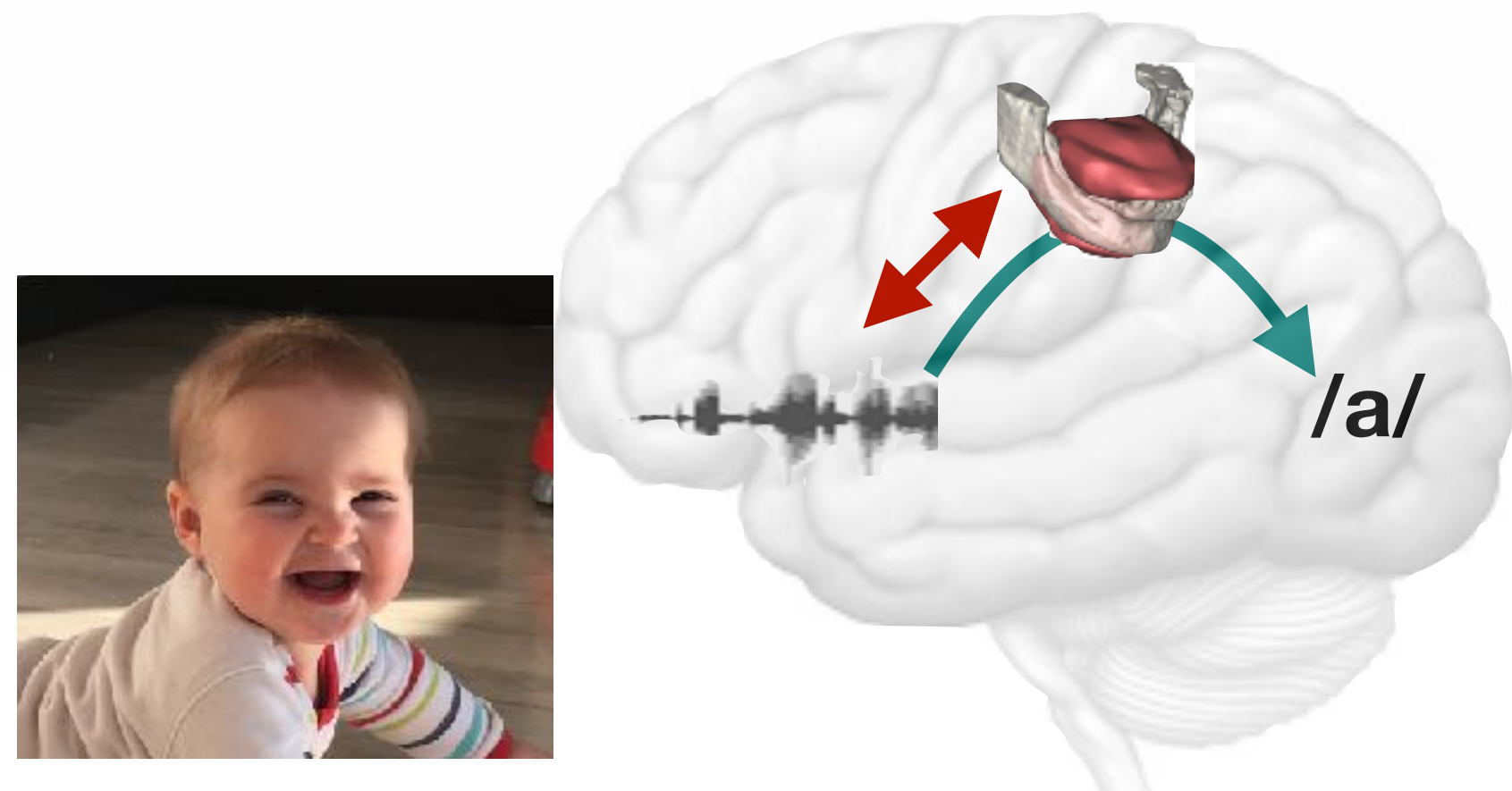
- An « inverse model » transforms an acoustic target into a set of motor commands (Tourville, Reilly, Guenther, 2008), (Houde et Nagarajan., 2011), (Perrier et al., 2012)

Acoustic-to-articulatory inverse mapping

- A ill-posed problem: non-linear & many-to-one (Atal, 1978), (Qui & Carreira-Perpiñán, 2007), (Neiberg et al, 2008)
 - Several vocal tract configurations can give almost the same spectrum (e.g. bite-block experiments)
 - Inter-speaker variability (idiosyncrasy)
 - Specific anatomy/morphology, e.g. particular shape of the vocal cavities (Ladefoged and Broadbent, 1957)
 - Speaker-specific articulatory strategy (Mokhtari et al., 2007), (Story et al. 2005, 2007)
 - An interaction of both, e.g. articulation influenced by the shape of the palate (Fuchs et al., 2008)



from (Douros et al., 2019) - ArtSpeechMRIfr

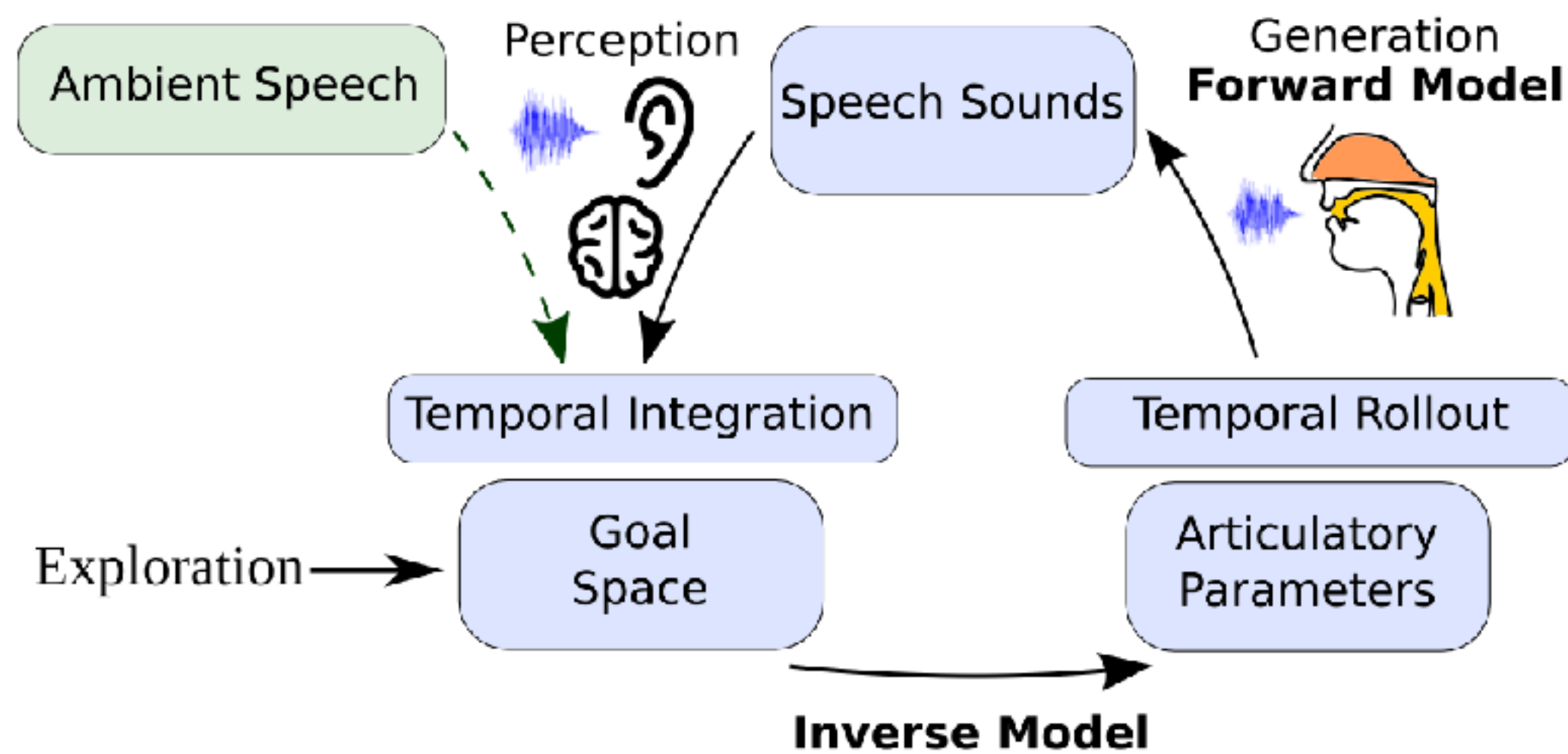


- Children seem to learn this complex inverse mapping mostly in a **self/weakly supervised manner (i.e. without supervisory feedback)**
 - for a given acoustic target, they are never provided with an explicit and complete feedback on the corresponding vocal tract configuration (our working hypothesis here...)

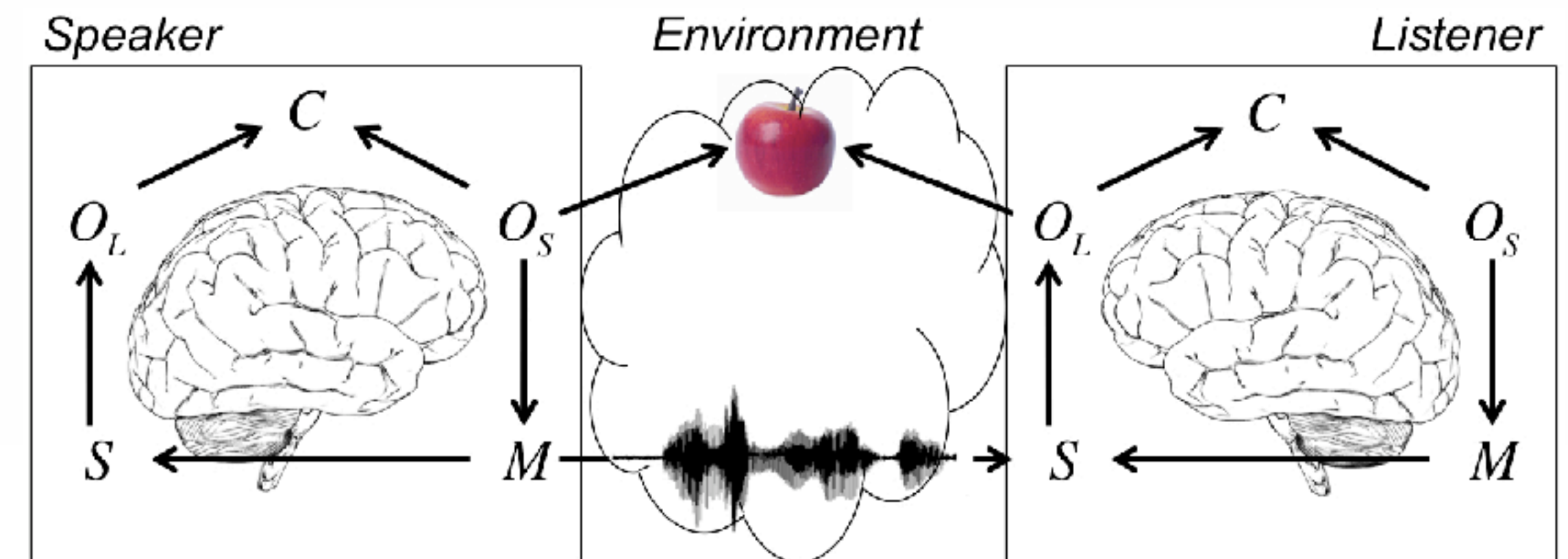
Computational model of speech learning

- Several studies proposed computational models of speech acquisition with a focus on the articulatory-acoustic mapping (Moulin-Frier et al., 2014), (Rasilo et Räsänen, 2017) (Philippsen et al., 2021), (Pitti et al., 2021)
- However, most models are evaluated on relatively simple linguistic material (isolated vowels, syllables) and/or synthetic speech data

Philippsen et al. 2021



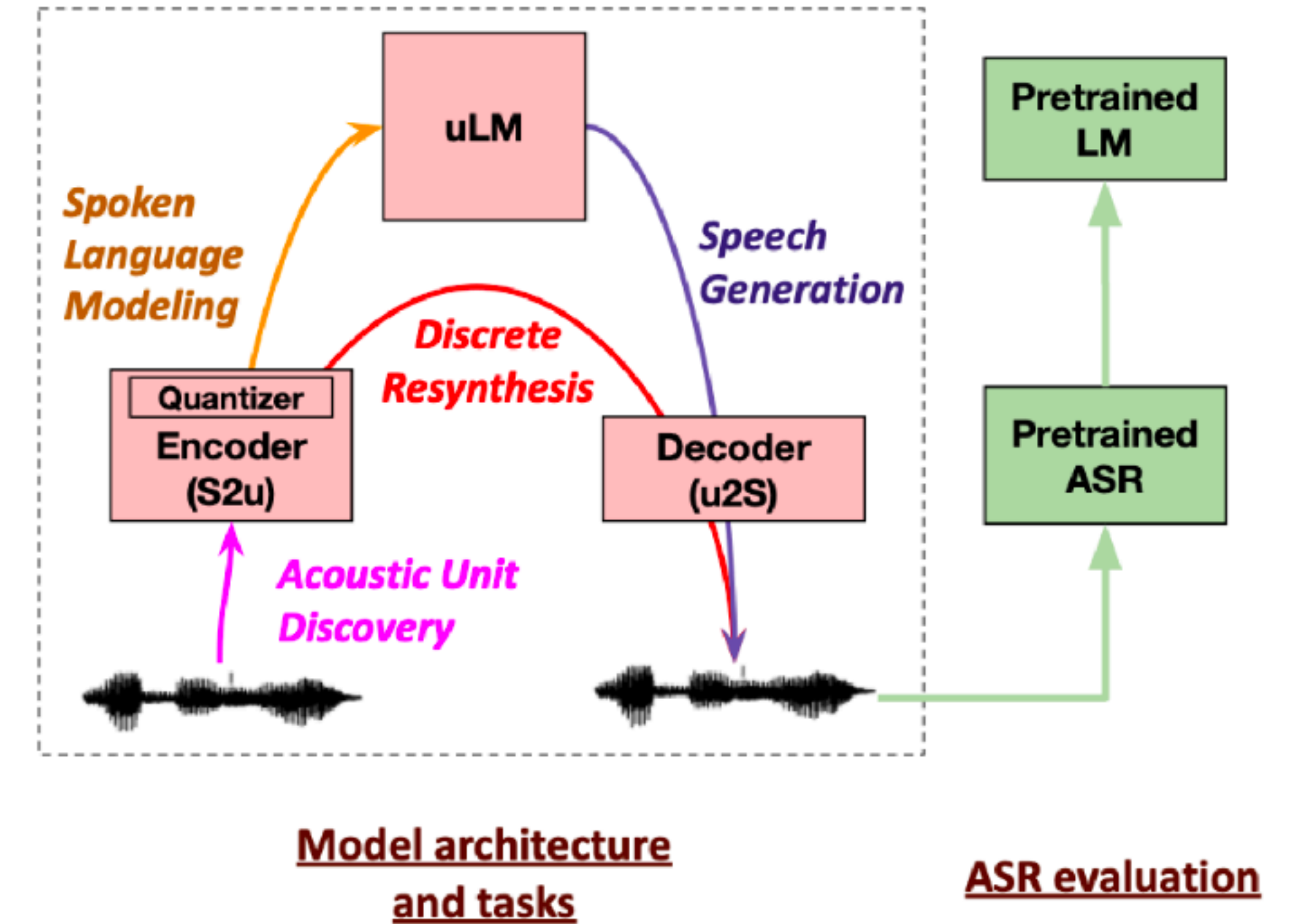
Moulin Frier et al, COSMO model (2011)



Deep learning approach

- Deep learning models neural networks, trained in a self-supervised manner, can be used as « tools » to study language and speech acquisition
 - Reverse engineering approach (Dupoux, 2016)
« constructing scalable computational systems that can, when fed with realistic input data, mimic language acquisition as it is observed in infants ».
 - Computational platform for testing which « innate learning constraints are necessary for speech and language acquisition » (Linzen, 2018)
- Able to deal with real-world data, can scale, can deal with raw data (limiting potential bias)
- Zero-resource challenges (Shatz et al., Dupoux et al.)
 - Learning the acoustic and linguistic characteristics of a language from raw audio (discovering Speech units, lexicon, discrete resynthesis, etc.)
- Articulatory level is almost never considered

Lakhotia et al, 2021,
On Generative Spoken Language Modeling from Raw Audio

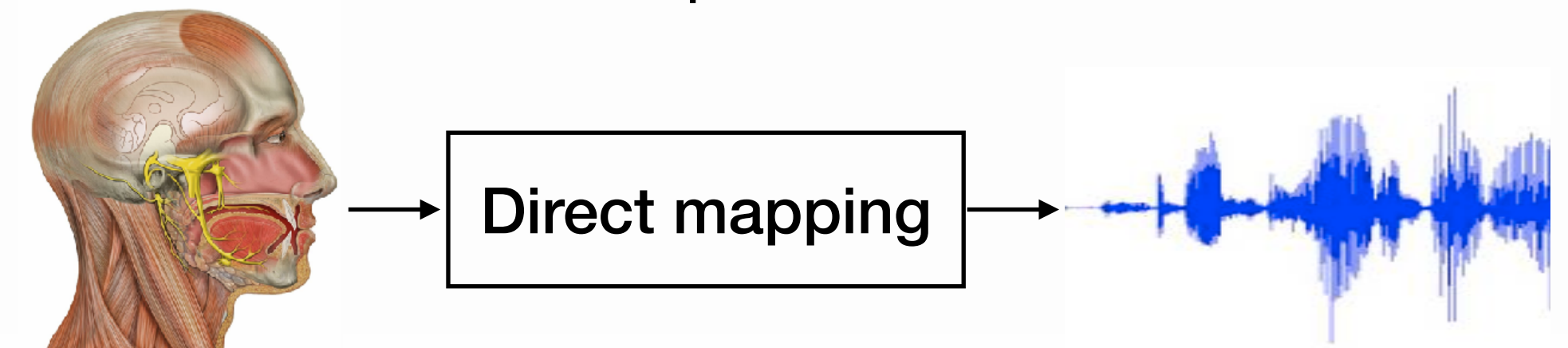


Long term goal: Build a computation model of speech acquisition, based on self-supervised deep learning, with explicit access to articulatory/motor knowledge

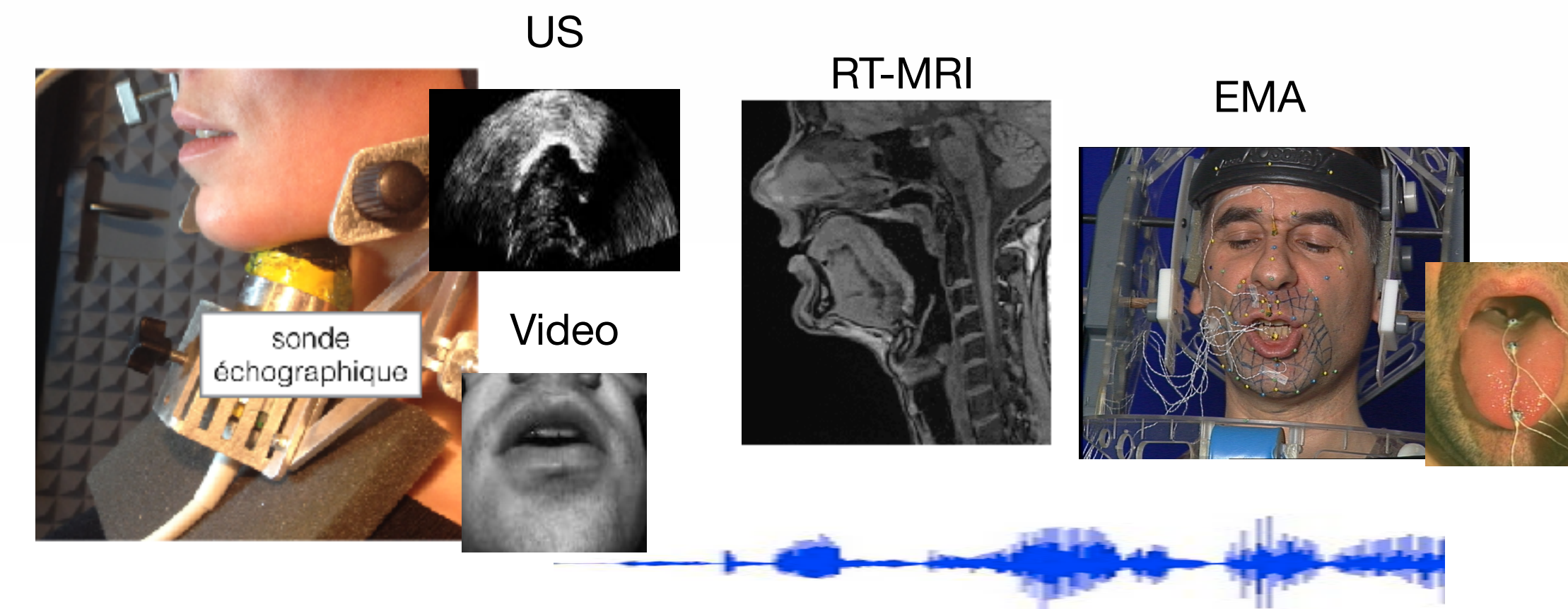
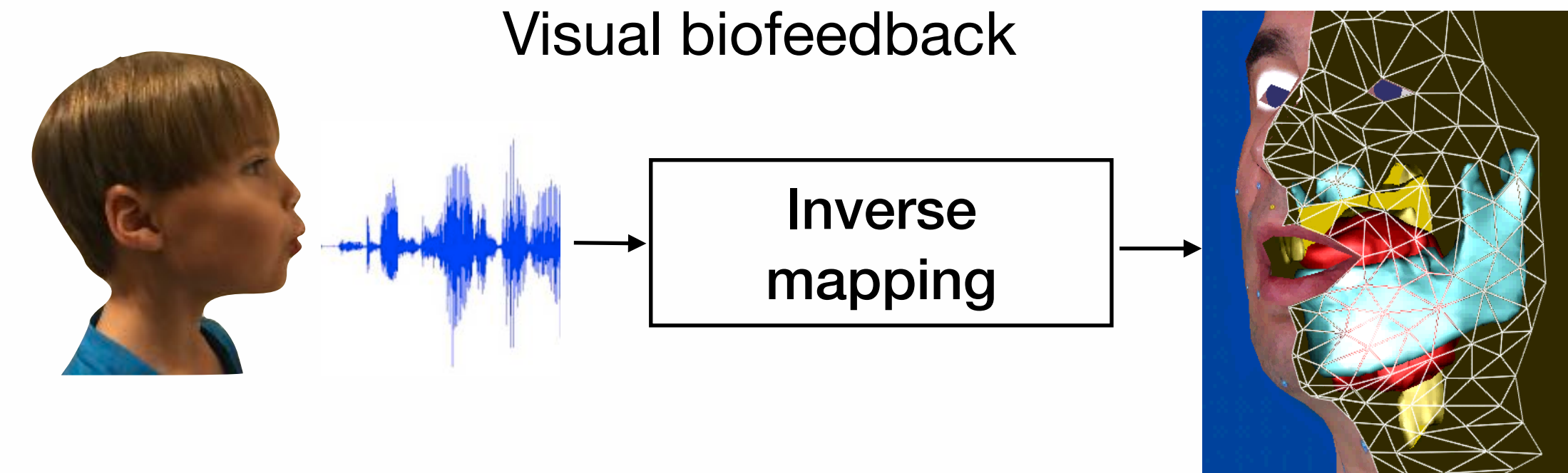
Articulatory synthesis & inversion

- A parallel line of research, with technological goals, e.g.
 - Direct mapping: articulatory synthesis / silent speech interface
 - Inverse mapping: visual biofeedback
- Direct or inverse models built from:
 - Parallel recordings of articulatory and acoustic speech data
 - Supervised training
 - Direct mapping: GMM (Toda et al., 2006), HMM (Hueber et al, 2006), DNN (Aryal & Gutierrez-Osuna, 2016), RNN (Taguchi & T. Kaburagi, 2018), Encoder-Decoder, (Chen et al. 2021)
 - Inverse mapping: Codebook (Ounie et Laprie, 2005), ANN (Richmond, 2004), GMM (Toda et al., 2008), HMM (Hiroya et al, 2004), DNN (Uria et al, 2011), RNN (Liu et al., 2015), Encoder-Decoder (Udupa, et al 2022)
 - Semi-supervised training / model adaptation
 - Inverse (Hueber et al., 2015), (Girin et al., 2016) & Direct mapping (Bocquelet et al., 2016)

Silent speech interfaces

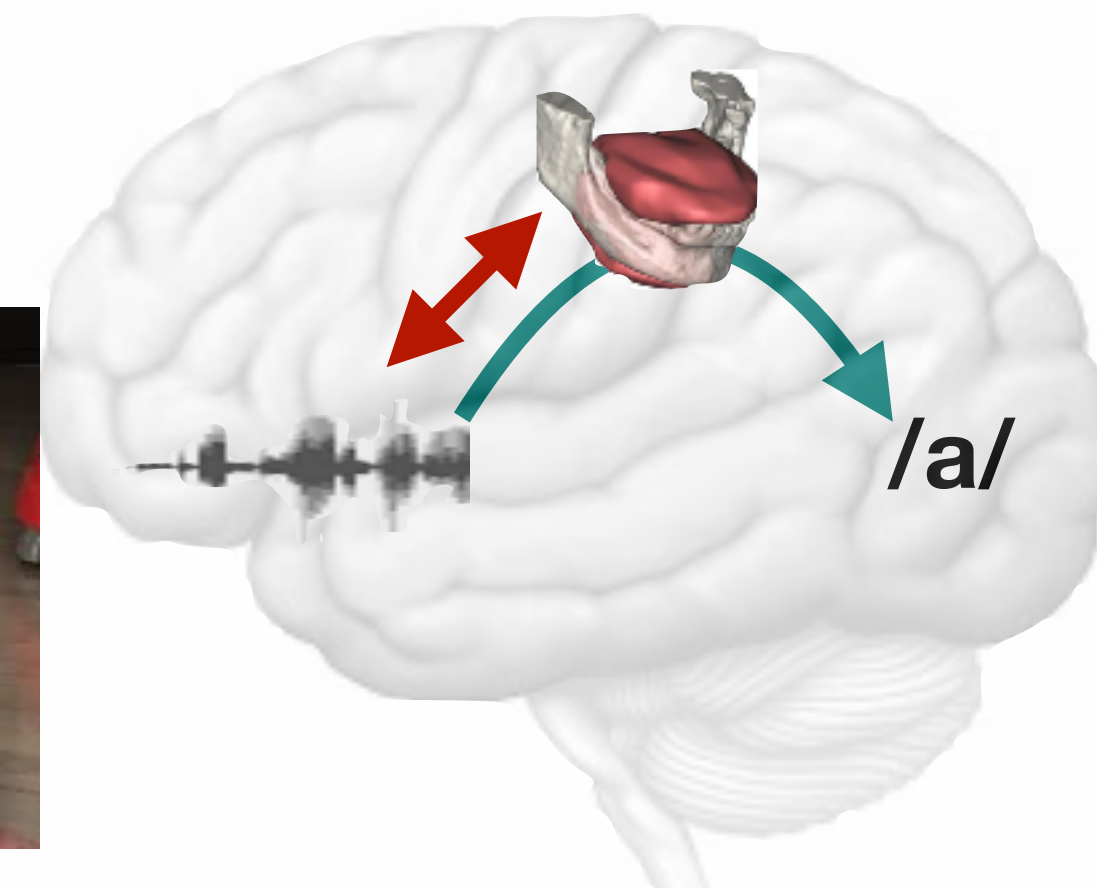


Visual biofeedback



Research goals and questions

- Goal: studying speech acquisition via self-supervised learning models with a focus on the role of motor/articulatory representations
- Research questions:
 - Does articulatory knowledge improve speech representations learned in a self-supervised manner? (Hueber et al. Neural Comp., 2020)
 - Does an explicit access to articulatory knowledge improve speech decoding ? (e.g. in adverse conditions) (Georges et al., Interspeech 2021)
 - How inverse acoustic-to-articulatory inverse mapping can be learned in a self-supervised manner? (Georges et al., ICASSP 2022)
 - What is the the role of articulatory knowledge in the discovery of phonological units? (Georges et al., Interspeech 2022)



Marc-Antoine Georges's PhD!

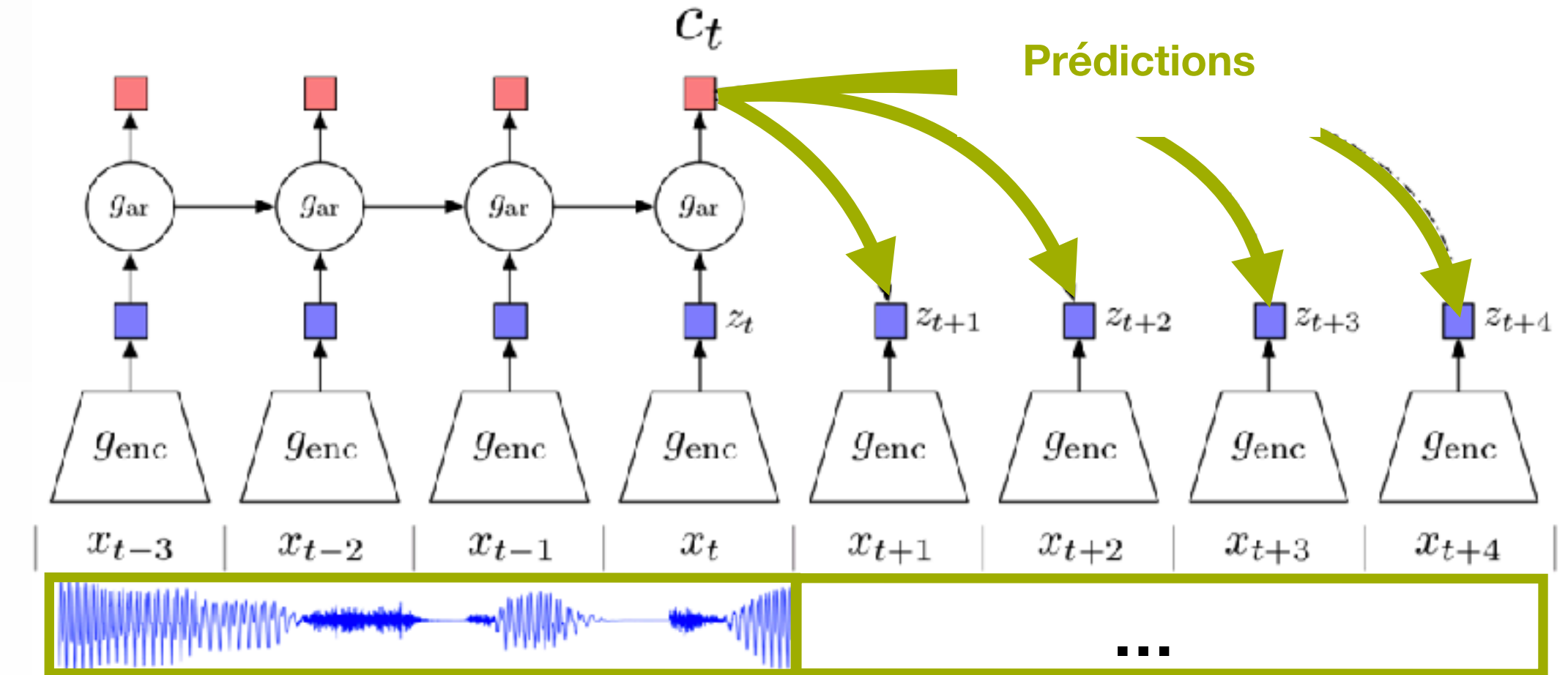
Predictive coding of audiovisual speech

(Hueber et al. Neural Computation, 2020)

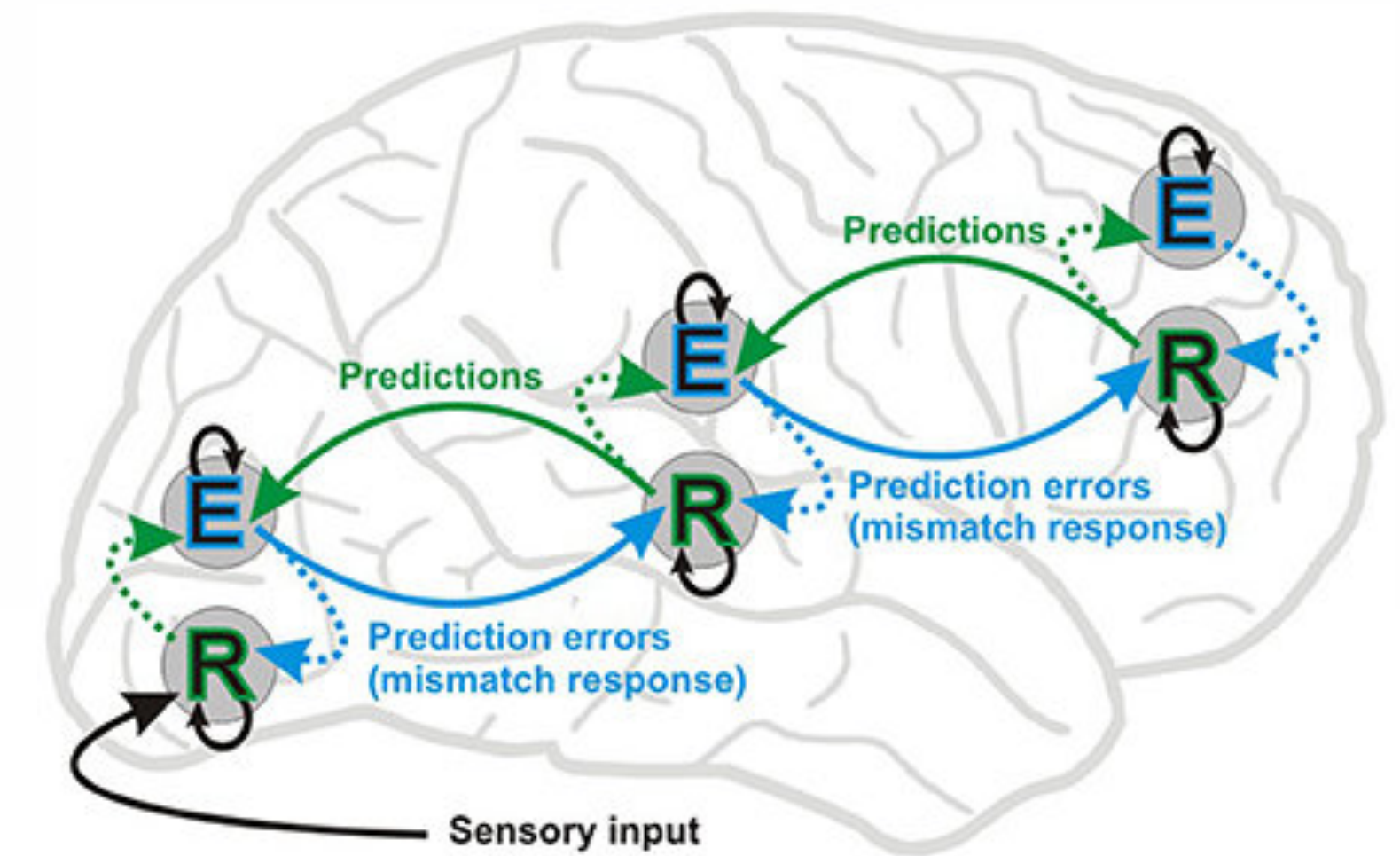
- Question: Does articulatory knowledge improve speech representation learned in a self-supervised manner?
- Predictive coding : a widely used pretext task in SSL of speech representation
 - Predicting the future the speech signal
(Chung et al., 2020 et al.), (van der Oord et al., 2019)
- A classic framework in cognitive & neuroscience (Rao, Ballard, 1999), (Friston 2003-2005-2011), (Hovsepyan et al., 2020)
 - **The brain is described as a 'prediction machine'** constantly forming predictions about upcoming input which guide the interpretation of sensory data
 - Goal: minimizing prediction error / surprise

Is there a benefit to exploit articulatory movements (mainly the lips) in addition to auditory speech ?

(Hypothesis: Lip movements anticipate the sound ...)



from van der Oord et al., 2019, CPC

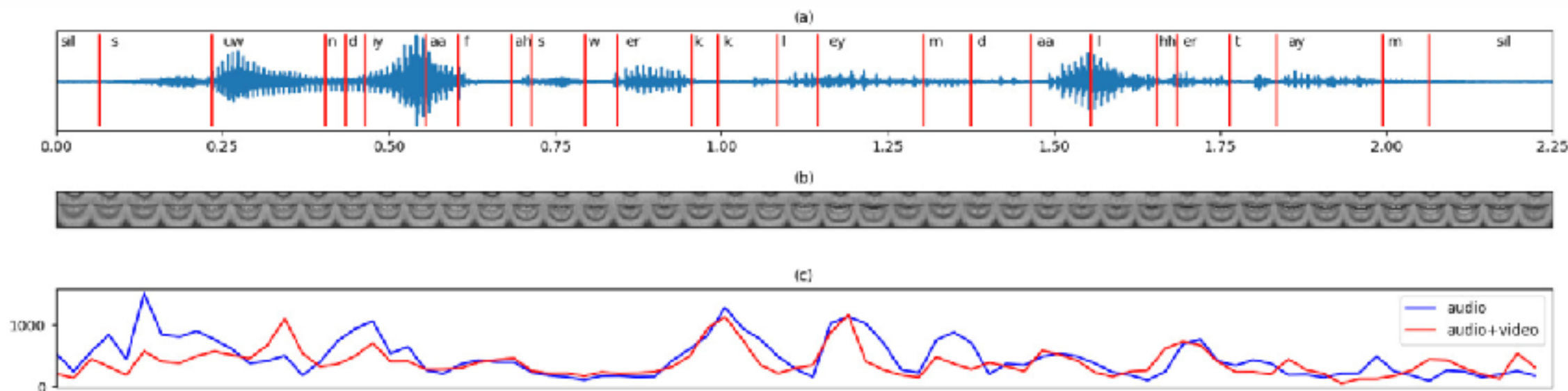
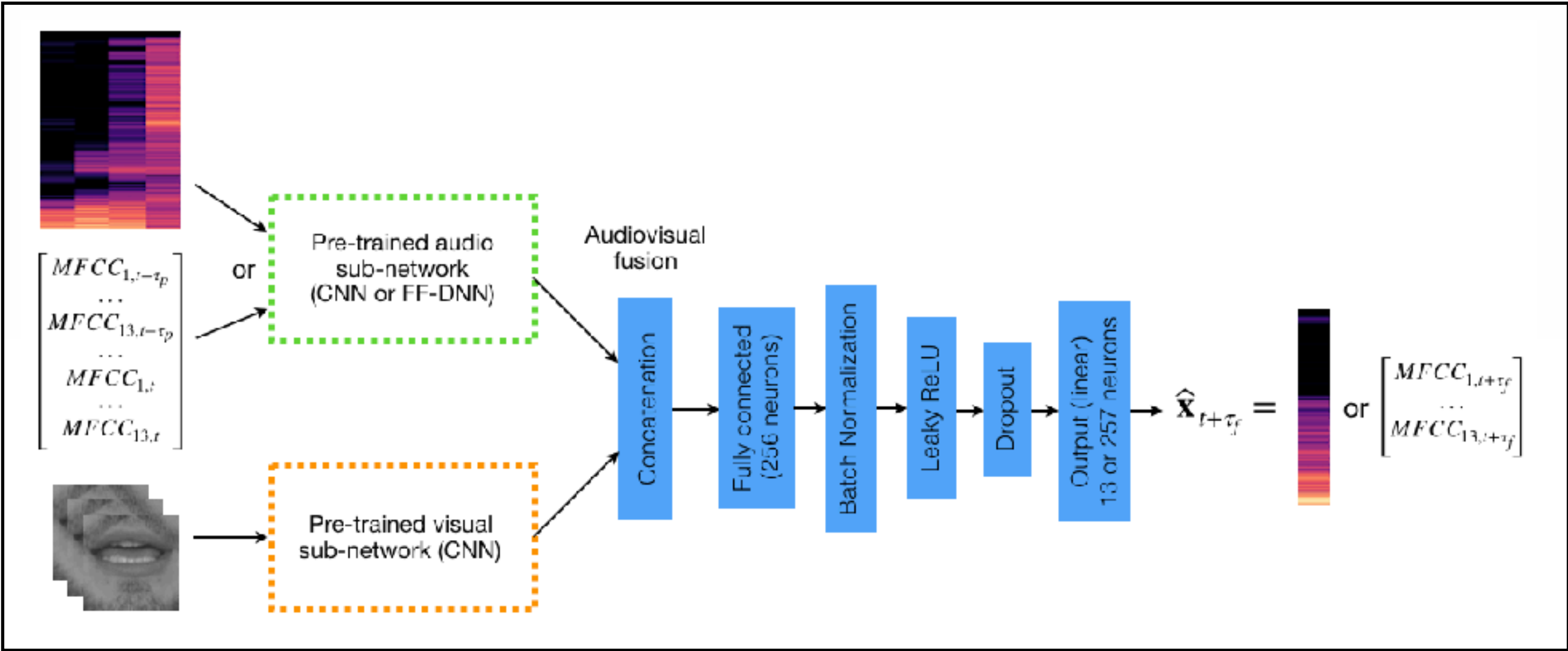
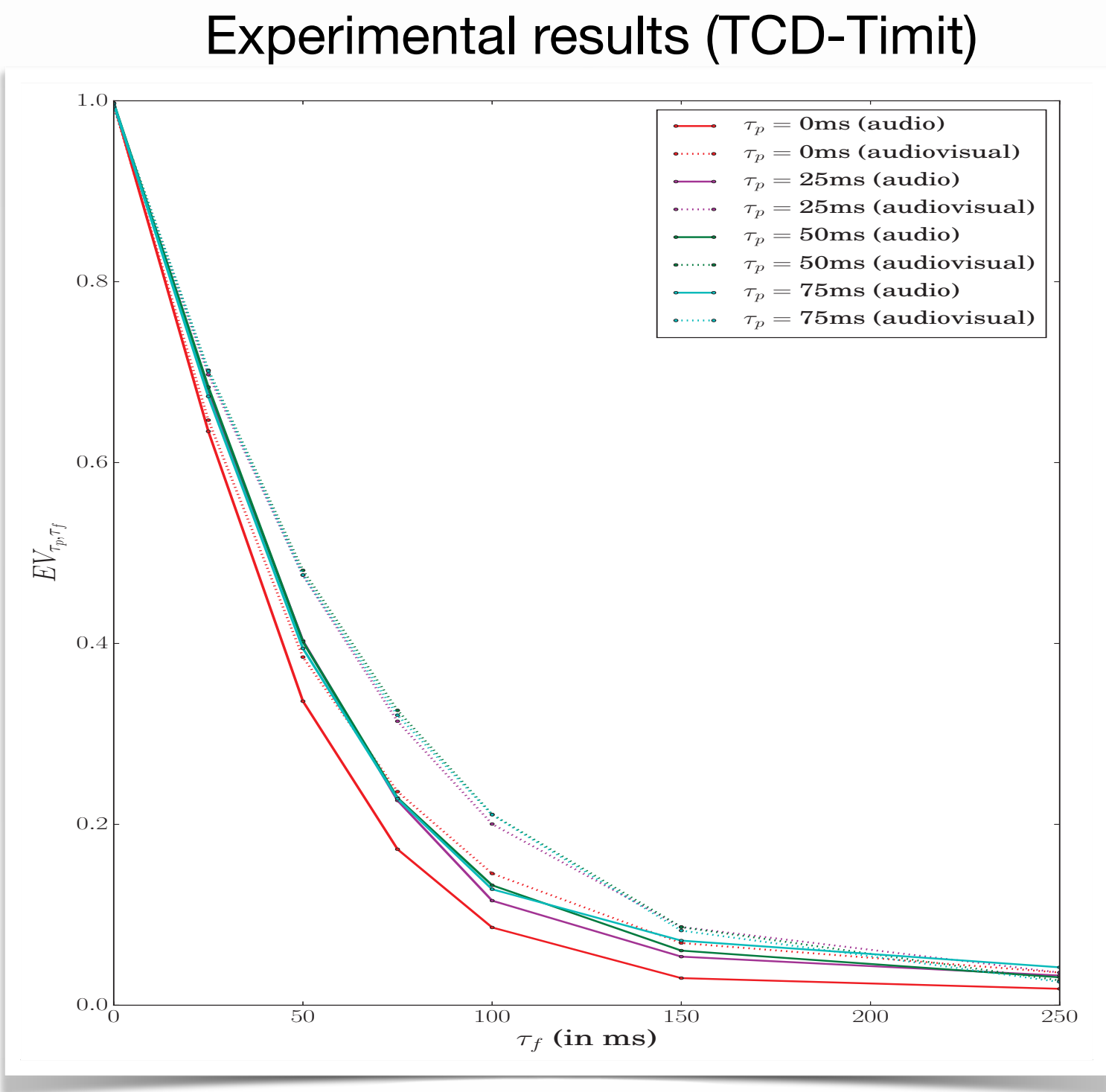
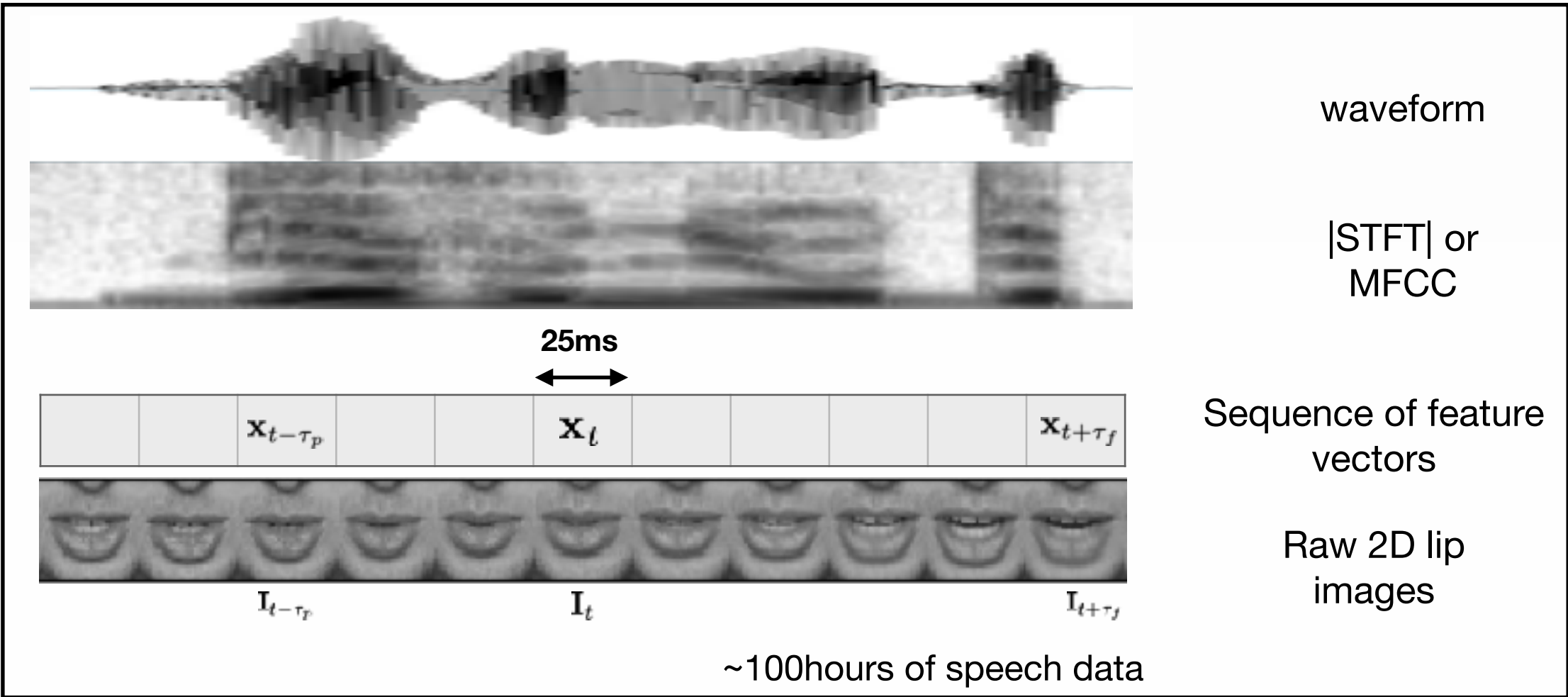


Simplified scheme of the hierarchical predictive coding framework (Friston, 2005, 2008, 2010)

Predictive coding of audiovisual speech

(Hueber et al. Neural Computation, 2020)

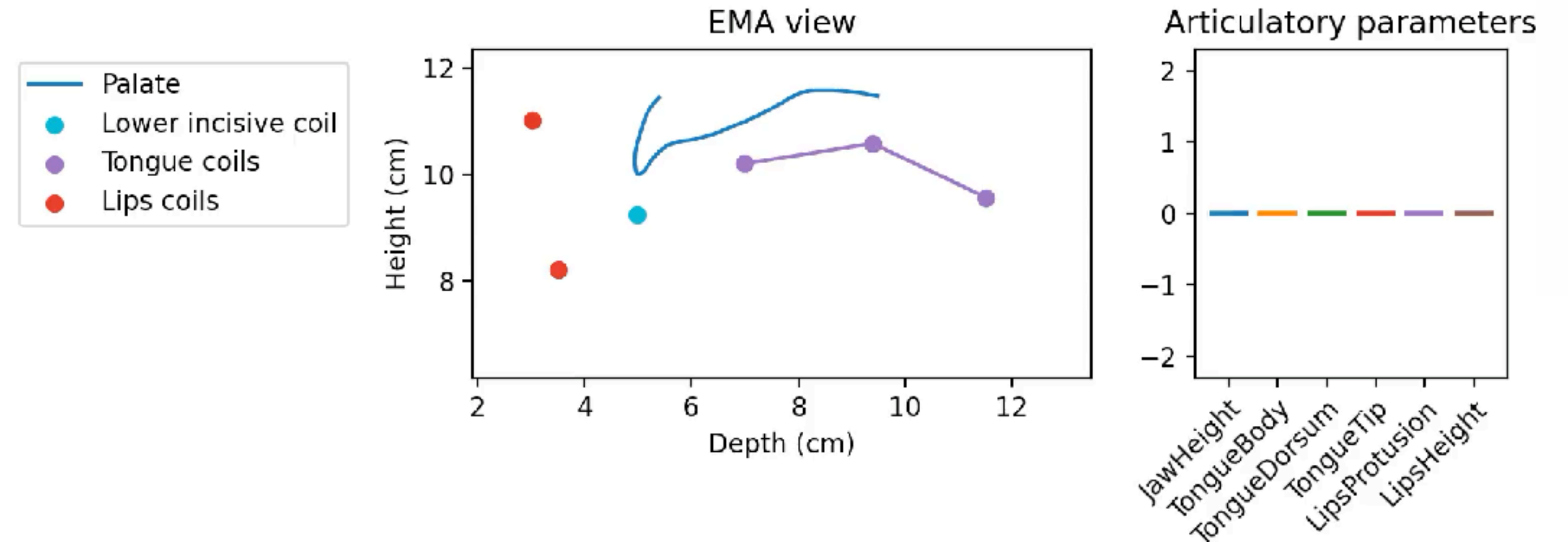
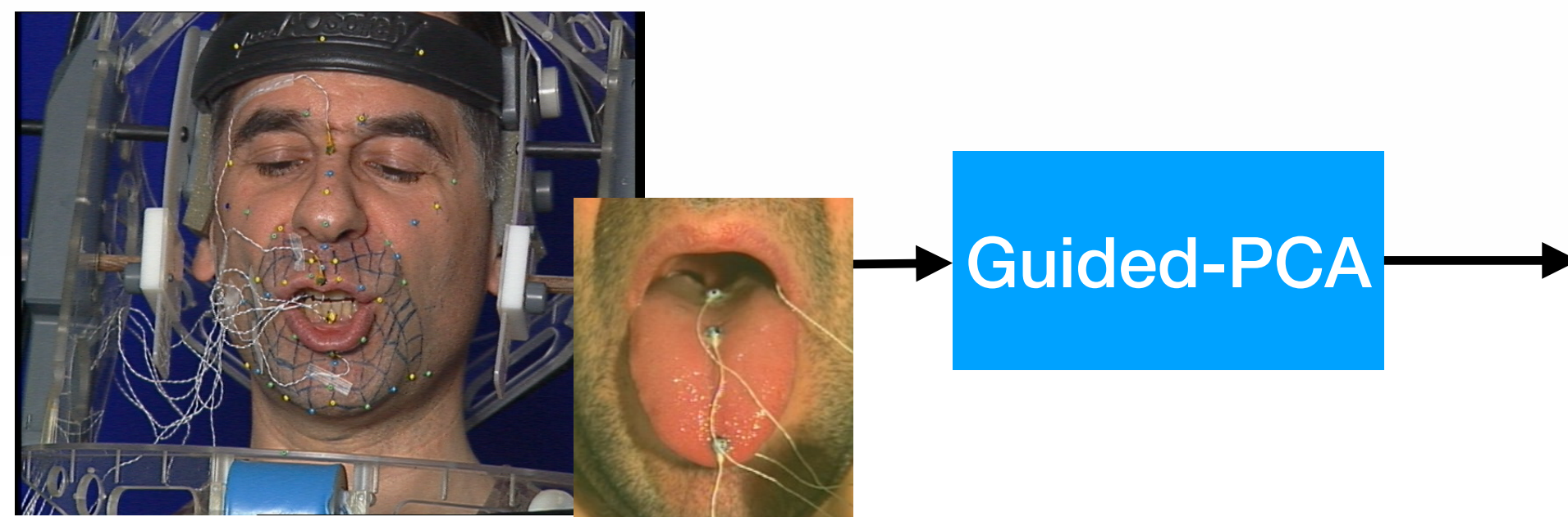
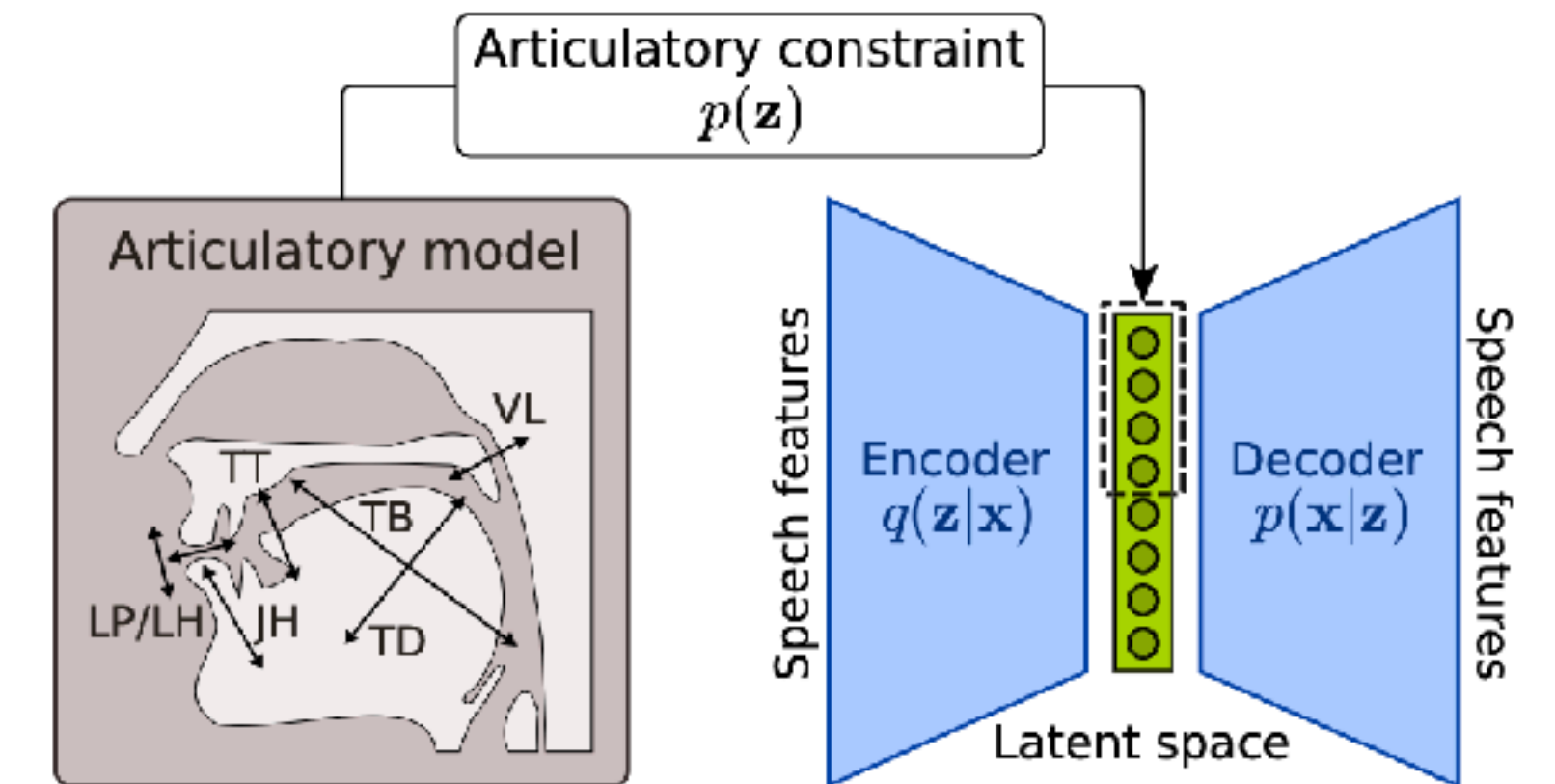
Predicting the future ... from the past auditory/visual inputs $\hat{\mathbf{x}}_{t+\tau_f} = f(\mathbf{x}_t, \mathbf{I}_t, \mathbf{x}_{t-1}, \mathbf{I}_{t-1}, \dots, \mathbf{x}_{t-\tau_p}, \mathbf{I}_{t-\tau_p})$



Articulatory-regularized VAE

(Georges et al, 2021)

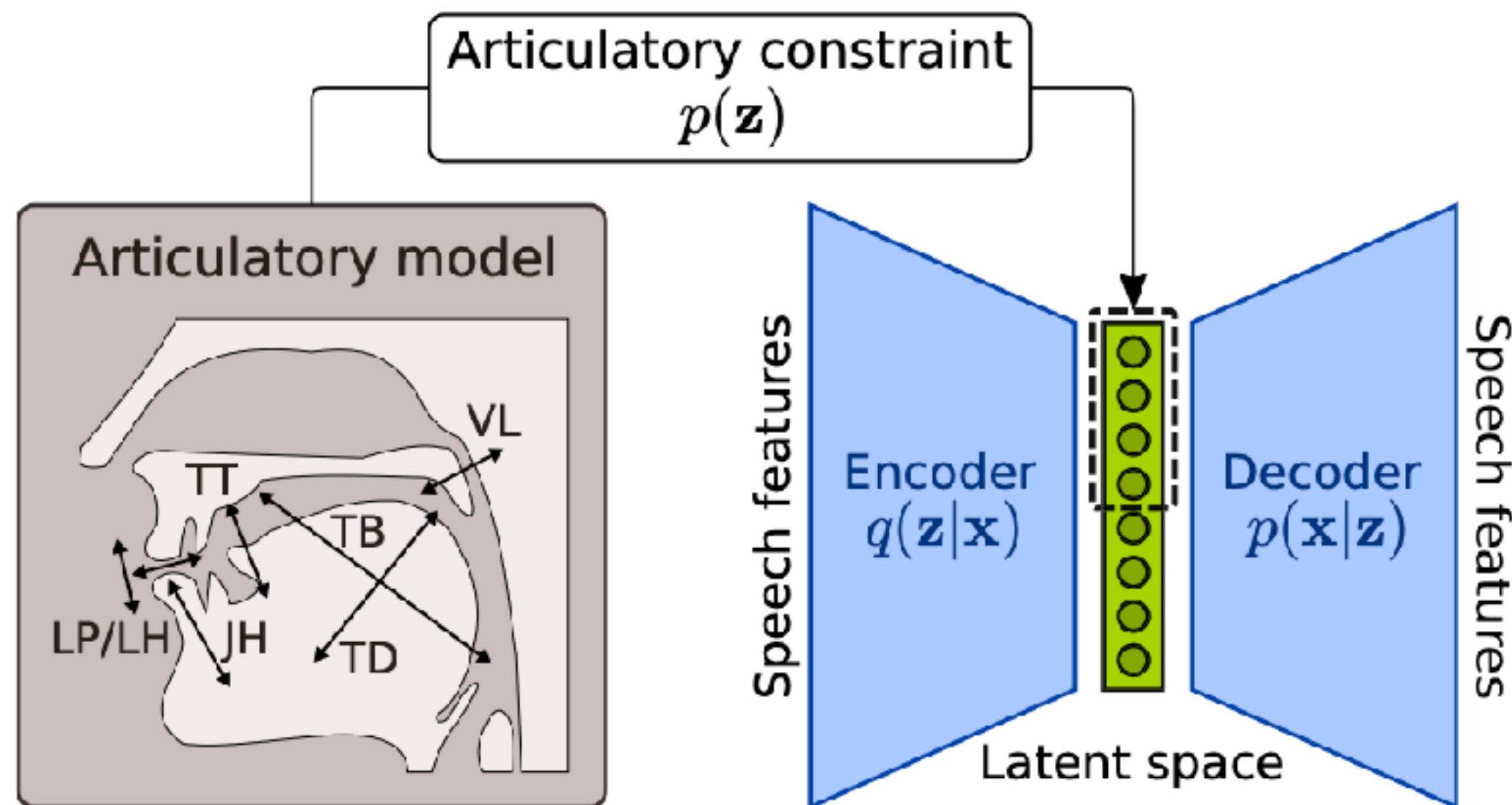
- Question: Does an explicit access to articulatory knowledge improve speech decoding in adverse conditions?
- Proposed approach: Speech denoiser based on VAE & articulatory knowledge in its latent space
- Step 1: Extract a linear articulatory model from raw EMA recordings from a reference speaker
 - Building an « articulatory model » using a simple decomposition technique called « guided-PCA », similarly to as (Maeda, 1979) (Beautemps et al., 2001) (Badin et al. 2002)



Articulatory-regularized VAE

(Georges et al, 2021)

- Step 2: Constraining some of the latent variables of the VAE so that they have the same distribution as these 7 articulatory parameters
- Initially proposed for controlling the timbre of music synthesizer sounds (Roche et al., TISMIR, 2021)



$$\mathcal{L}(\phi, \theta, \mathbf{x}) = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\log p_{\theta}(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}[q_{\phi}(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})] + \alpha \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\mathcal{R}(\mathbf{z}, \mathbf{a}(\mathbf{x}))].$$

with

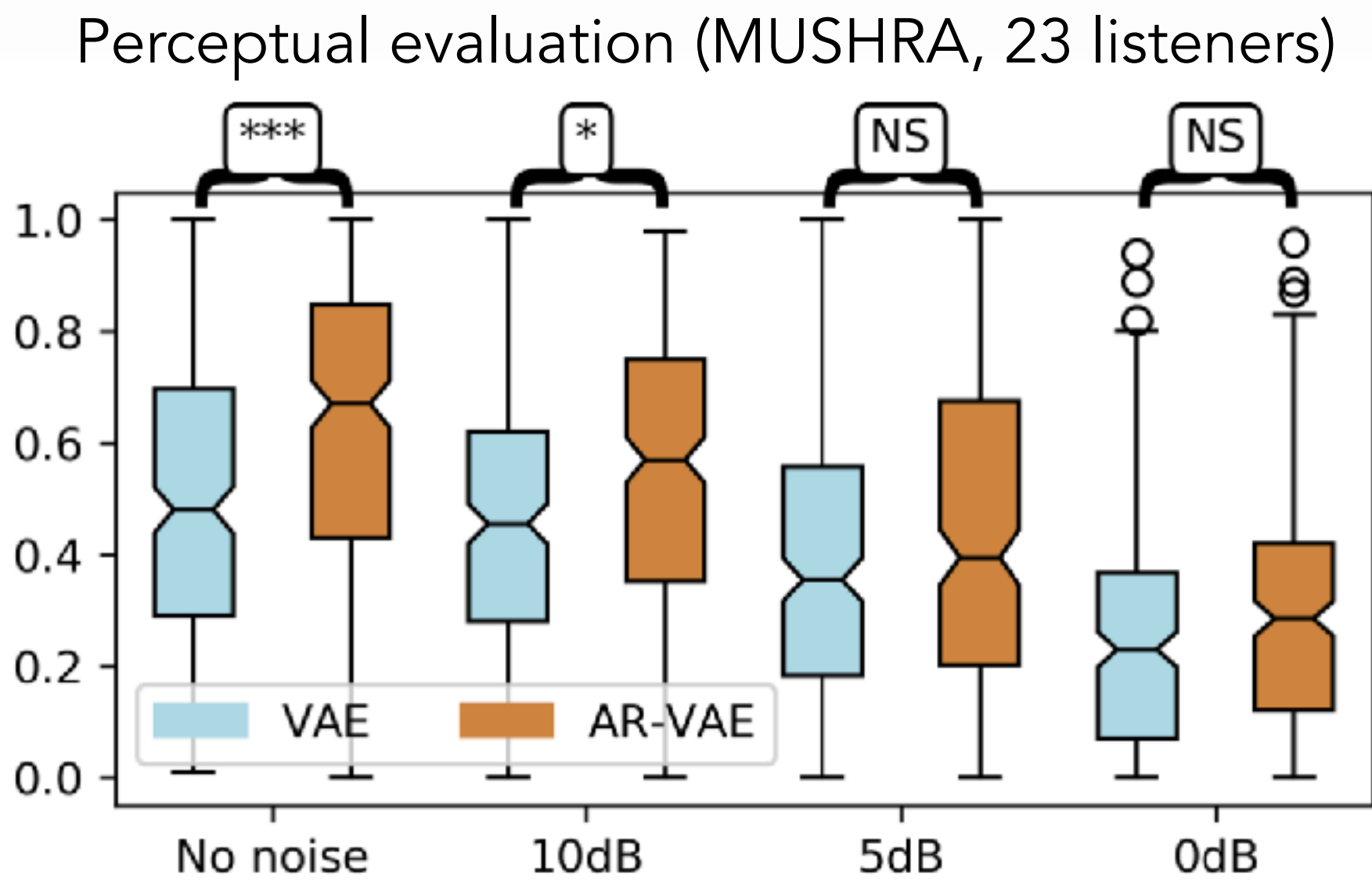
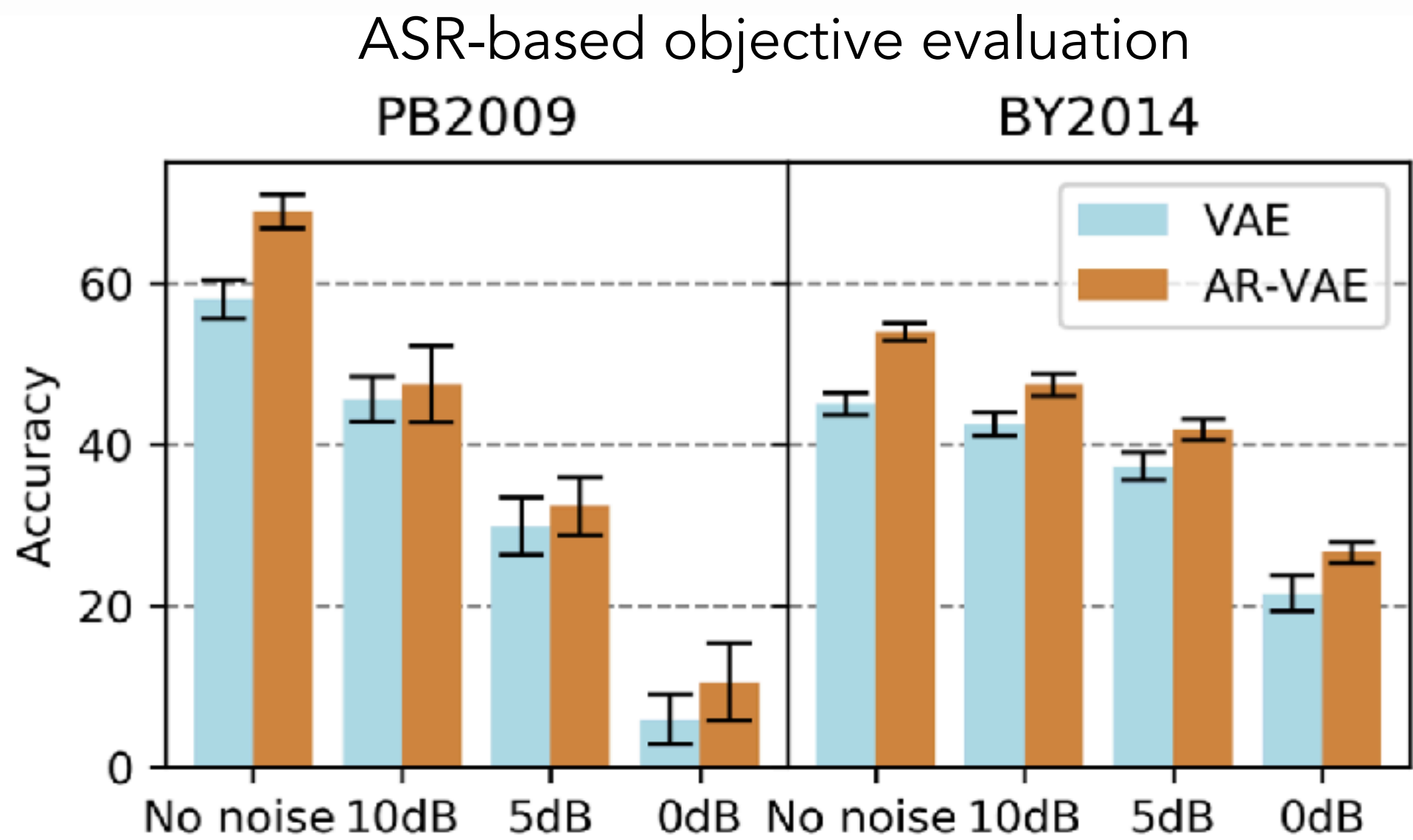
$$\mathcal{R}(\mathbf{z}, \mathbf{a}(\mathbf{x})) = \|\mathbf{z}_{1:N} - \mathbf{a}(\mathbf{x})\|^2.$$

↑
articulatory parameters
associated with the acoustic observation \mathbf{x}

Articulatory-regularized VAE

(Georges et al, 2021)

Implementation details: speaker-dependent models, encoder&decoders implemented as 3 layers DNN with 100 neurons in each layer, speech signal encoded using 18 Bark-scale coefficients and reconstructed using LPCNet neural vocoder






- AR-VAE slightly outperforms VAE, but no massive effect

Noisy speech VAE AR-VAE

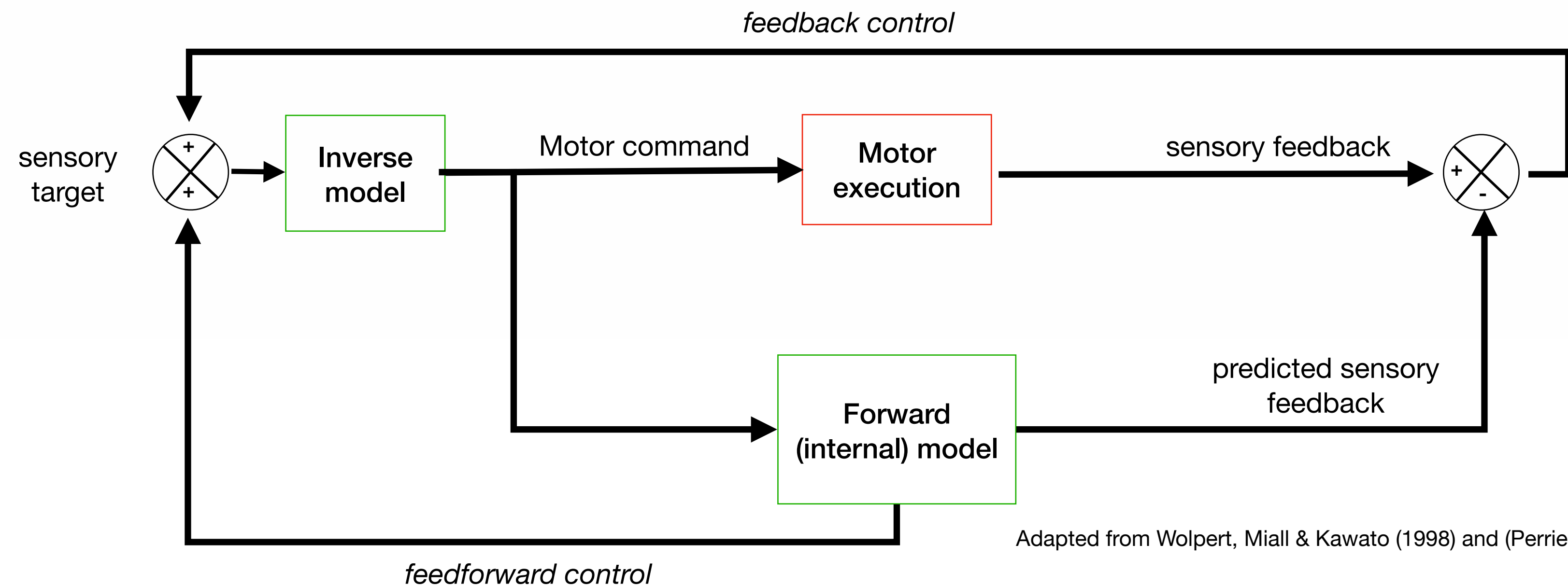
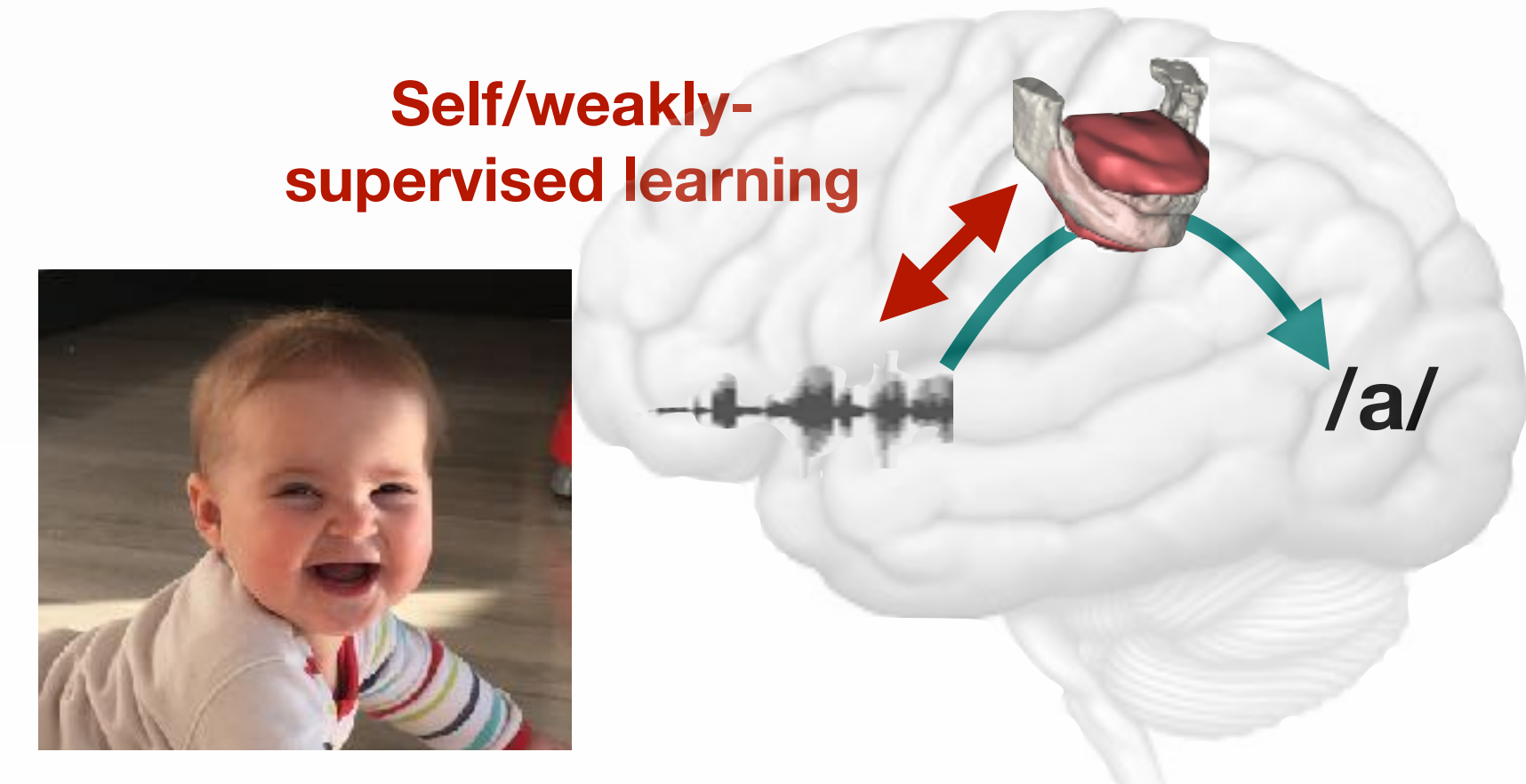
Noisy speech VAE AR-VAE

SSL of acoustic-to-articulatory inverse mapping

(Georges et al., 2021-2022)

- Question: How a child can learn inverse the acoustic-to-articulatory mapping in a self-supervised manner?
- Computational model of motor control
(Wolpert, Miall & Kawato (1998), (Tourville, Reilly, Guenther, 2008), (Houde et Nagarajan., 2011), (Perrier 2012)
 - feedback control (early stage of learning, adverse condition) —> slow
 - feedforward control (« automatic pilot ») —> fast

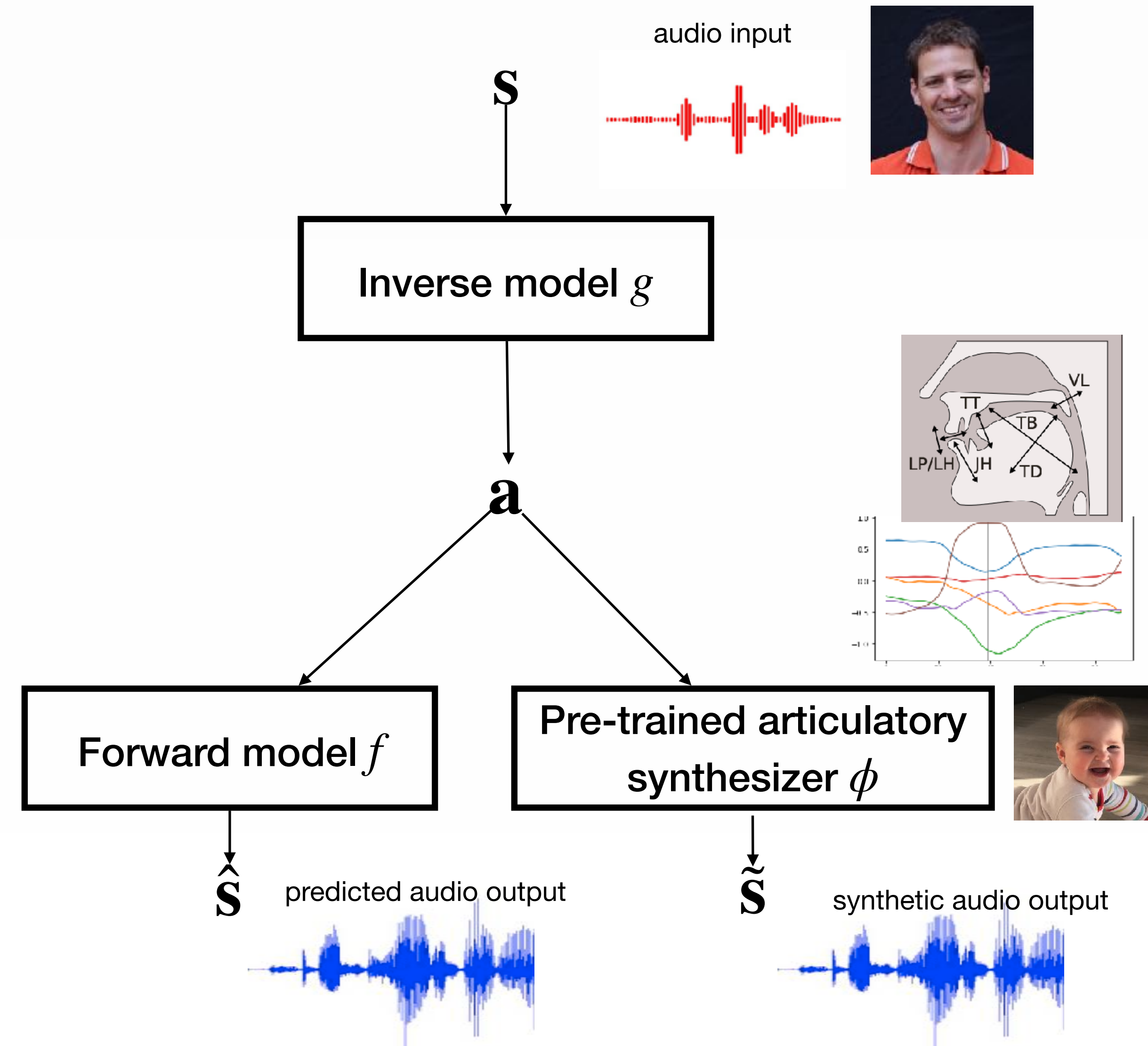


Adapted from Wolpert, Miall & Kawato (1998) and (Perrier 2012)

SSL of acoustic-to-articulatory inverse mapping

(Georges et al., 2021-2022)

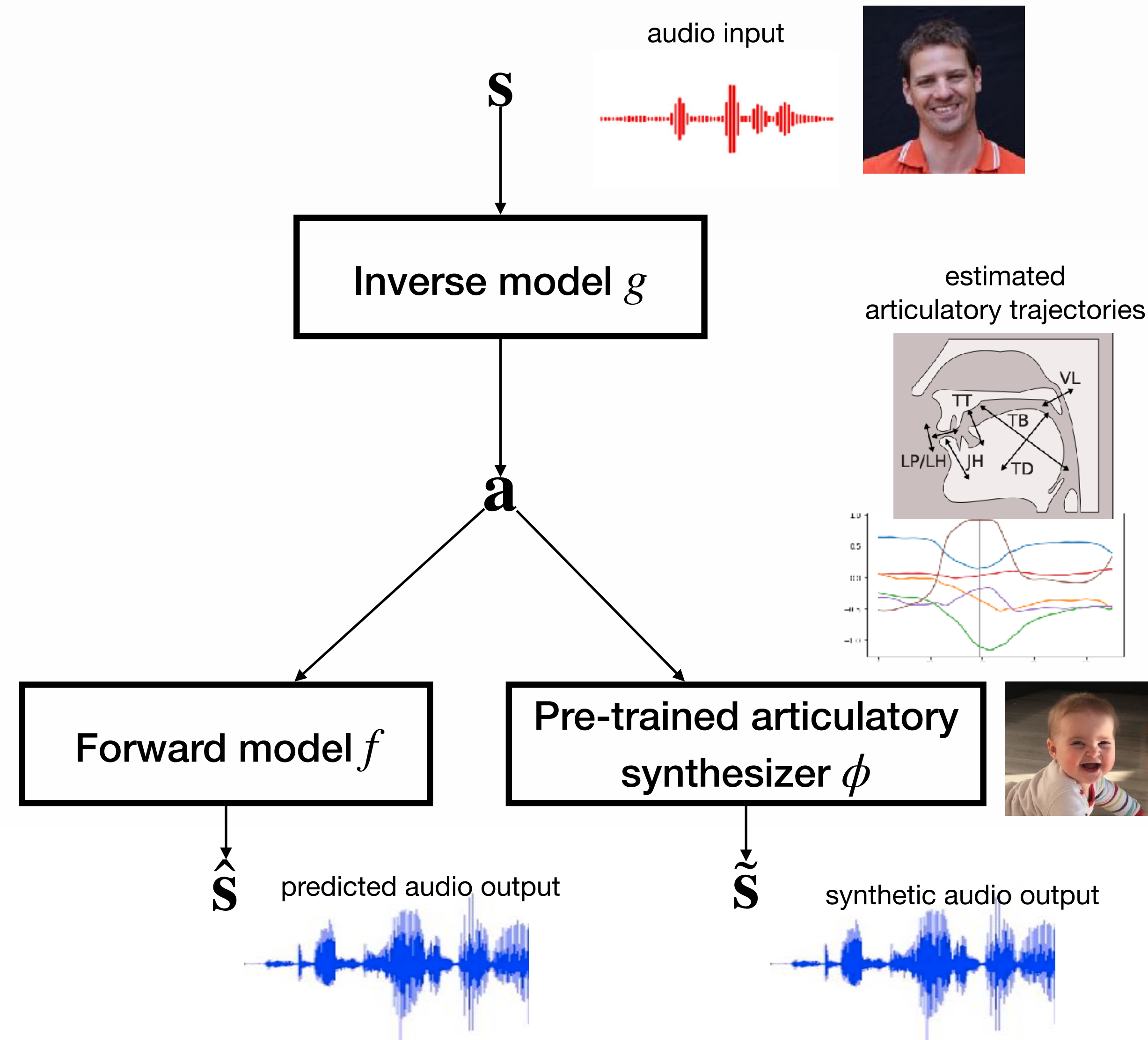
- Computational model of speech production based on DNN
 - Pre-trained (and frozen) articulatory synthesizer
 - Input: 7 articulatory parameters
 - Output: 18-dimensional Bark-scale coefficients
 - Pretrained on a EMA-audio recordings (~30mn, 1 male speaker)
 - Architecture 4 layers of 200 neurons, standard training procedure
 - Audio synthesis using a neural vocoder (LPCNet) + original f0
 - Inverse model recovering articulatory parameters from the acoustic speech input
 - unidirectional LSTM, 2 layer, 32 units each
 - Forward model predicting the acoustic consequences of articulatory commands
 - same architecture as the synthesizer, but trainable



SSL of acoustic-to-articulatory inverse mapping

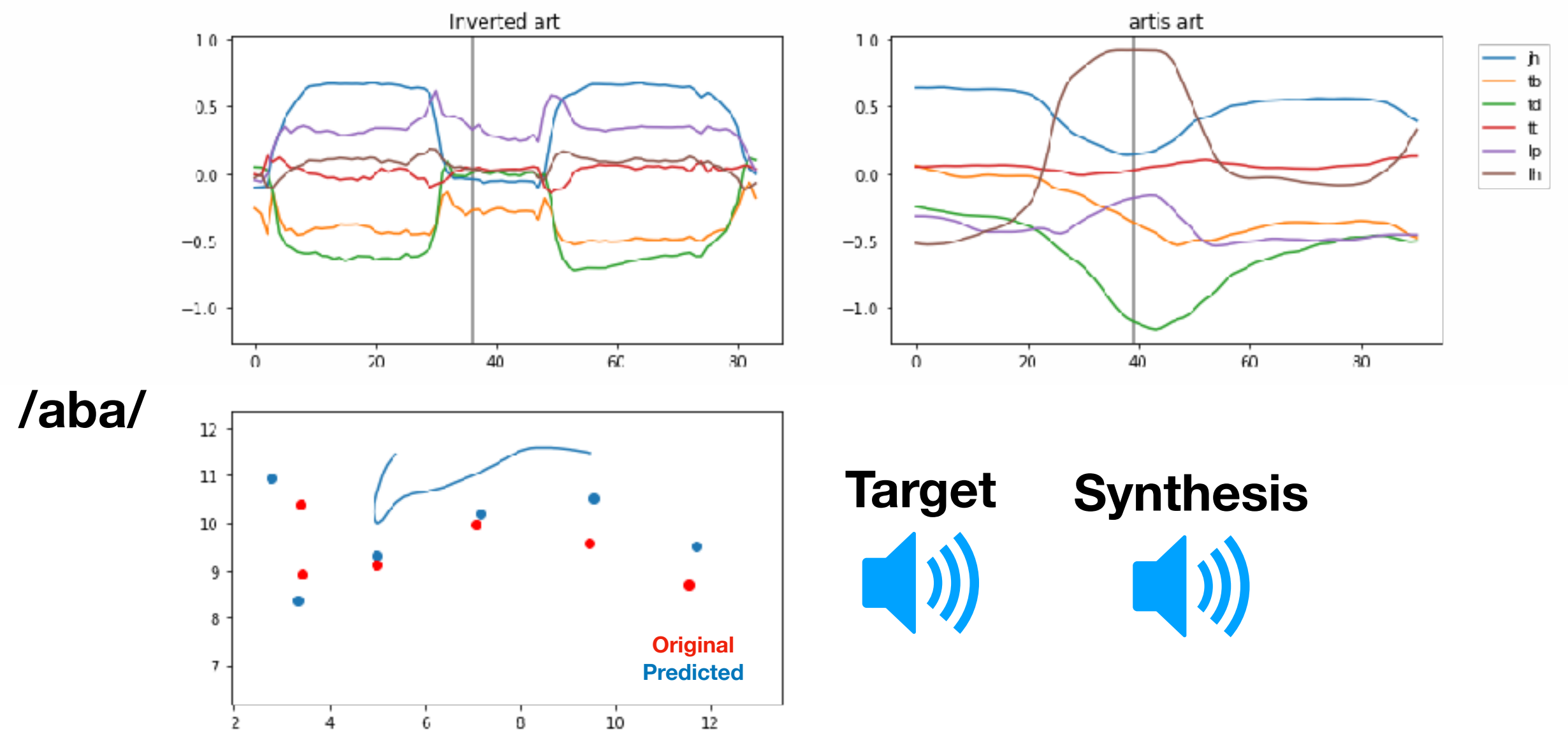
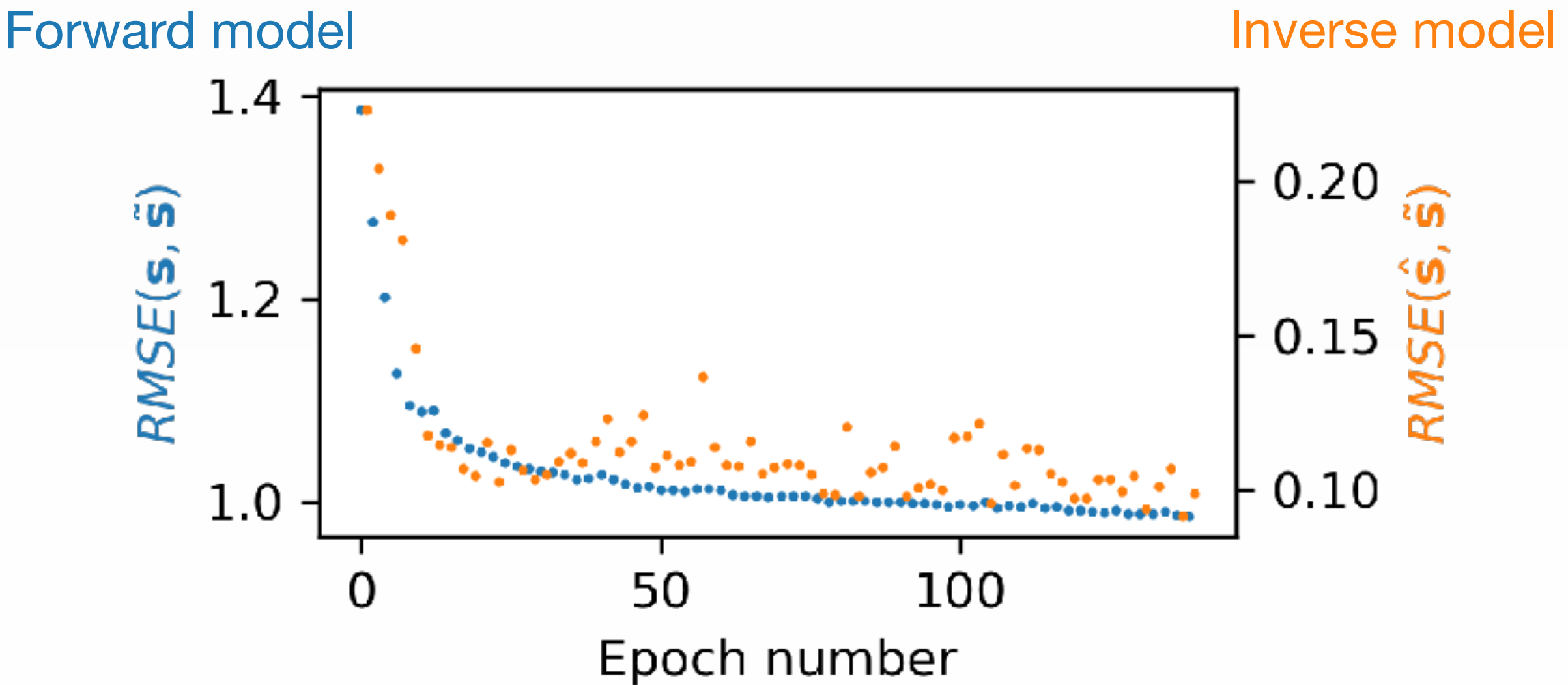
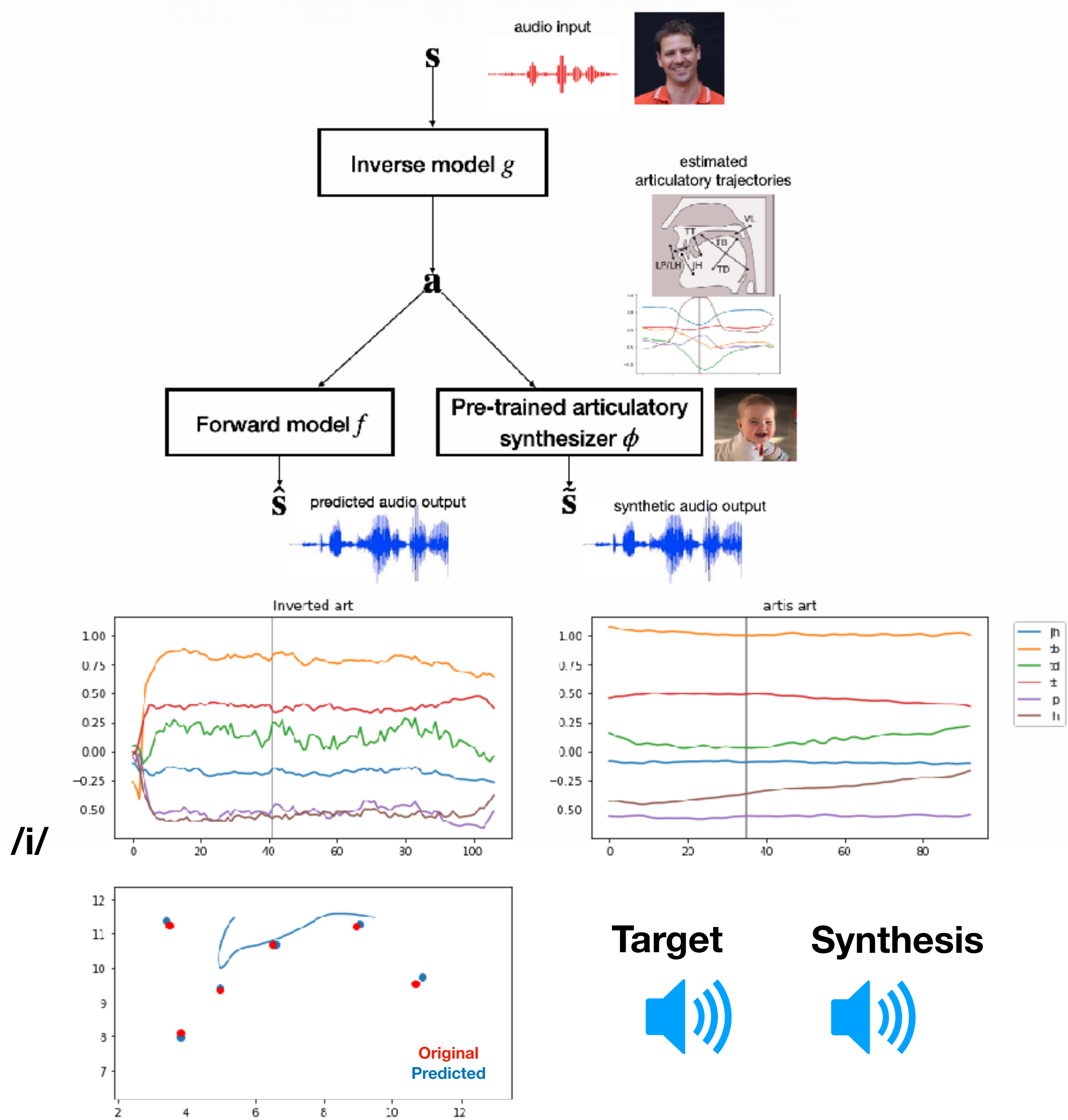
(Georges et al., 2021-2022)

- Model trained end-to-end to repeat audio stimuli from 3 other speakers
- Starting from a random initialization
 - Inverse mapping $\mathbf{a} = g(\mathbf{s})$
 - Articulatory synthesis $\tilde{\mathbf{s}} = \phi(\mathbf{a})$ and forward mapping $\hat{\mathbf{s}} = f(\mathbf{a})$
 - Update of the forward model (backprop.)
 - $L_{int} = \|\tilde{\mathbf{s}} - \hat{\mathbf{s}}\|$ ability of the forward model to approximate the physical system (i.e. the synthesizer)
 - Update the inverse model (backprop. but with the forward model frozen)
 - $L_{ext} = \|\tilde{\mathbf{s}} - \mathbf{s}\|$ discrepancy between audio input and synthetic audio output



SSL of acoustic-to-articulatory inverse mapping

(Georges et al., 2021-2022)



Conclusions and perspectives

- Self-supervised deep learning model can be used to study speech learning mechanisms
 - with a focus on acoustic-articulatory mapping
- SSL of speech representation from multimodal input (lips + audio) using predictive coding
 - Information provided by the visual modality is real but limited.
 - Visual-only information does not evidence a stable advance of lips on sound (as sometimes stated in the literature).
- Articulatory-regularized VAE
 - Outperforms VAE but by a small margin only ... (we need to scale here ...)
 - Speech perception: a small (but significant) benefit of relying on articulatory prior knowledge for decoding speech in adverse conditions
- Computation model of speech acquisition inspired by speech motor control model
 - Self-supervised learning of acoustic-to-articulatory inverse model
- Current work / perspectives:
 - Introducing biomechanical constraints in the inverse model
 - Investigating the role of articulatory representations for the discovery of phonological units (see Georges et al., Interspeech 2022)
 - Connecting this research with developmental & neuro-physiological data (holy grail?) ...

The end!

Thanks!

Collaborative work:

Eric Tatulli (post-doc, now at TIMC)
Marc-Antoine Georges (PhD, GIPSA-lab, LPNC)
Fanny Roche (Arturia, GIPSA-lab)
Laurent Girin (GIPSA-lab)
Jean-Luc Schwartz (GIPSA-lab)
Julien Diard (LPNC)
Xavier Alameda-Pineda (INRIA, Montbonnot)
Pierre Badin (GIPSA-lab)

