Why audio-visual learning improves voice identity recognition: a neurocomputational model

Christian Gumbsch (christian.gumbsch@tu-dresden.de)
Chair of Cognitive and Clinical Neuroscience, Faculty of Psychology, TU Dresden
& Neuro-Cognitive Modeling, Faculty of Science, University of Tübingen

Martin V. Butz (martin.butz@uni-tuebingen.de)
Neuro-Cognitive Modeling, Faculty of Science, University of Tübingen,
Tübingen, Germany

Katharina von Kriegstein (katharina.von_kriegstein@tu-dresden.de)
Chair of Cognitive and Clinical Neuroscience, Faculty of Psychology, TU Dresden
Dresden, Germany

Abstract

Voice identity recognition in auditory-only conditions is facilitated by knowing the face of the speaker. This effect is called the 'face-benefit'. Based on neuroscience findings, we hypothesized that this benefit emerges from two factors: First, a generative world model integrates information from multiple senses to better predict the sensory dynamics. Second, the model substitutes absent sensory information, e.g., facial dynamics, with internal simulations. We have developed a deep generative model that learns to simulate such multisensory dynamics, developing latent speaker-characteristic contexts. We trained our model on synthetic audio-visual data of talking faces and tested its ability to recognize speakers from their voice only. We found that the model recognizes previously seen speakers better than previously unseen speakers when given their voice only. The modeling results confirm that multisensory simulations and predictive substitutions of missing visual inputs result in the face-benefit.

Keywords: multisensory learning, speech, voice, world models

Introduction

Human communication is multimodal. For example, hearing and seeing a speaker talk improves speech understanding (Peelle & Sommers, 2015). Perhaps surprisingly, visual information about a speaker's face can also help process their speech later on when only auditory speech is provided: many experiments have shown that when participants first saw the face of a speaker talking, they were later better at identifying this speaker or recognizing spoken words under auditory-only conditions, compared to a control condition where the face was not shown (von Kriegstein et al., 2008; Maguinness, Schall, & Kriegstein, 2021). Based on neuro-imagining studies, von Kriegstein et al. (2008) hypothesized that this so-called **'face-benefit'** develops from internal simulations of familiar speakers' faces. The face-benefit occurs both for auditory-only voice identity recognition as well as for auditory-only speech recognition. So far, the face-benefit and neuroimaging findings have only been explained qualitatively.

Here, we present the first approach to model the face-benefit for voice identity recognition. Our main contributions are the following:

- We present a multimodal world model that jointly encodes and simulates signals from different sensory modalities.
- We show that face-benefits develop from internal face simulations of familiar speakers, providing the first model-based explanation for the findings in humans.

Methods & Material

Multimodal World Model

In line with theories of predictive processing (Friston, 2010; Huang & Rao, 2011; Hohwy, 2013), we hypothesized that humans maintain an internal generative model that continuously attempts to predict future sensory input. We implement such a world model (Ha & Schmidhuber, 2018; Friston et al., 2021) as a Recurrent State Space Model (Rssm, Hafner et al., 2019). Importantly, we extend the Rssm by (1.) **multimodal processing pathways** and (2.) **cross-modal simulations**.

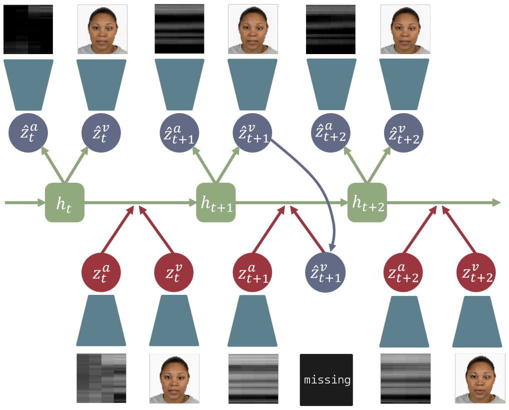

Figure 1: **Multimodal world model** that learns to predict auditory-visual inputs by developing modality-specific codes (z_t^a, z_t^v) and supramodal hidden states ($\boldsymbol{h}_t$). Via internal simulation (purple arrow) missing inputs are substituted.

Our model with trainable parameters ϕ is computed by:[1]

$$\text{Full state } \boldsymbol{s}_t \leftarrow (\boldsymbol{h}_t, \boldsymbol{z}_t^a, \boldsymbol{z}_t^v) \quad (1) \qquad \text{Dynamics } \boldsymbol{h}_t = f_\phi(\boldsymbol{s}_{t-1}) \quad (5)$$

$$\text{Audio prior } \hat{\boldsymbol{z}}_t^a = p_\phi^a(\boldsymbol{h}_t) \quad (2) \qquad \text{Vis. prior } \hat{\boldsymbol{z}}_t^v = p_\phi^v(\boldsymbol{h}_t) \quad (6)$$

$$\text{Audio post. } \boldsymbol{z}_t^a = q_\phi^a(\boldsymbol{h}_t, \boldsymbol{o}_t^a) \quad (3) \qquad \text{Vis. post. } \boldsymbol{z}_t^v = q_\phi^v(\boldsymbol{h}_t, \boldsymbol{o}_t^v) \quad (7)$$

$$\text{Audio dec. } \hat{\boldsymbol{o}}_t^a = d_\phi^a(\boldsymbol{h}_t, \boldsymbol{z}_t^a) \quad (4) \qquad \text{Vis. dec. } \hat{\boldsymbol{o}}_t^v = d_\phi^v(\boldsymbol{h}_t, \boldsymbol{z}_t^v) \quad (8)$$

The state $\boldsymbol{s}_t$ of the model (Eq. 1) is composed of a supramodal hidden state $\boldsymbol{h}_t$ and modality-specific sensory codes $\boldsymbol{z}_t^a$ and $\boldsymbol{z}_t^v$ for audio (a) and vision (v). At each time t, the model observes a new audio signal ($\boldsymbol{o}_t^a$) and an image ($\boldsymbol{o}_t^v$) and embeds these observations into its stochastic[2] sensory codes $\boldsymbol{z}_t^a$ and $\boldsymbol{z}_t^v$ (Eq. 3, Eq. 7). Subsequently, it updates its hidden state $\boldsymbol{h}_t$ (Eq. 5). From the new latent state, the model can reconstruct the observations $\boldsymbol{o}_t^a$ and $\boldsymbol{o}_t^v$ (Eq. 4, Eq. 8). Furthermore, the model makes prior predictions about its next sensory codes $\hat{\boldsymbol{z}}_t^a$ and $\hat{\boldsymbol{z}}_t^v$ (Eq. 2, Eq. 6), before new inputs arrive and the process is repeated (see Fig. 1).

To deal with partially missing observations, we implement a **cross-modal simulation mechanism**. When a certain observation $\boldsymbol{o}_t^i$ for modality $i \in \{v, a\}$ is missing, the model substitutes the sensory code $\boldsymbol{z}_t^i$ with its prior prediction $\hat{\boldsymbol{z}}_t^i$ (see Fig. 1). During model training, we randomly activate cross-modal simulation for visual inputs with 50% probability.

All components are implemented as neural networks, as in Hafner et al. (2020), and their parameters ϕ are trained to minimize a **variational free energy loss** $\mathcal{L}$, i.e., a computational neuroscience-inspired objective (Friston, 2010):

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi}\left[\sum_{i \in a,v} \mathcal{L}^{\text{NLL}}(\boldsymbol{o}_t^i, \hat{\boldsymbol{o}}_t^i, \phi) + \mathcal{L}^{\text{KL}}(q_\phi^i, p_\phi^i, \phi)\right] \quad (9)$$

[1] Blue components are newly added to the Rssm.

[2] Following Hafner, Lillicrap, Norouzi, and Ba (2020), we sample $\boldsymbol{z}_t^i$ from a vector of categorical distributions.

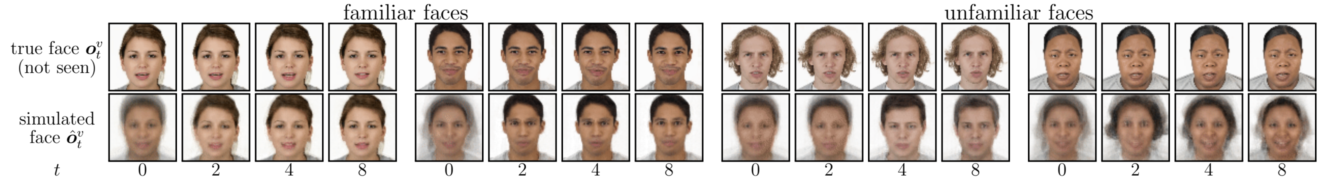
familiar faces
unfamiliar faces
true face $\boldsymbol{o}_t^v$ (not seen)
simulated face $\hat{\boldsymbol{o}}_t^v$
t
0
2
4
8
0
2
4
8
0
2
4
8
0
2
4
8

Figure 2: **The model simulates faces based on auditory-only speech input** for speakers whose face was seen during training (left) or not (right).

where the negative log-likelihood $\mathcal{L}^{\text{NLL}}$ strives for accurate reconstructions $\hat{o}_t^i$ and $\mathcal{L}^{\text{KL}}$ minimizes the KL divergence between prior p_ϕ^i and posterior q_ϕ^i of sensory codes $\boldsymbol{z}_t^i$ $(i \in a, v)$.

Speaker Classifier

To model experiments on voice identity recognition (Maguinness et al., 2021), we need a module to identify the speakers' identity. Decision making is implemented as a simplified drift-diffusion race model (Ratcliff & McKoon, 2008; Gold & Shadlen, 2007). We train a neural network with learnable parameters θ, which receives the state $\boldsymbol{s}_t$ of the world model as input and outputs a categorical probability $P_\theta(S^j \mid \boldsymbol{s}_t)$ over speakers S^j. At time t, a speaker choice is sampled $\hat{S}_t^j$ and internally aggregated:

Drift: $\hat{S}_t^j \sim P_\theta(S^j \mid \boldsymbol{s}_t)$ (10) Evidence: $S_{\text{sum}}^j \leftarrow S_{\text{sum}}^j + \hat{S}_t^j$ (11)

A decision is made as soon as the aggregated evidence S_{sum}^j for some j exceeds a threshold, i.e., $S_{\text{sum}}^j \geq 3$. We add a constant offset (530ms) as 'non-decision time' (Ratcliff & McKoon, 2008). Crucially, when visual input is missing, the network receives the simulated visual code $\hat{\boldsymbol{z}}_t^v$ as part of its input, i.e., $\hat{\boldsymbol{s}}_t \leftarrow (\boldsymbol{h}_t, \boldsymbol{z}_t^a, \hat{\boldsymbol{z}}_t^v)$. The network is trained to minimize the cross-entropy between true speakers S_t^j and predicted speakers $\hat{S}_t^j$.

Talking Faces Dataset

Humans learn to process speech from thousands of conversations over ontogenetic development. To simulate this, our model is trained on an audio-visual dataset of speakers. For data generation, we selected faces from the Chicago Face database (Ma, Correll, & Wittenbrink, 2015), a large-scale database of face images taken under controlled conditions. We assigned each face a synthetic voice of a text-to-speech program (pytssx3, Bhat, 2020). Voice parameters were partially inferred from image data (biological sex, timbre) or randomly assigned (pitch, speed, English dialect). We animated the faces with SadTalker (Zhang et al., 2023), a state-of-the-art model for animating faces from still images and audio. We generated sentences composed of number words $(0-30)$.

Videos were provided to the model as images (64×64 pixels, 25Hz). The audio was processed as a Mel spectrogram (32 channels, 220 hop length) and segmented over time into 25 non-overlapping spectrograms per second, such that each image of the video was accompanied by a separate spectrogram image (scaled up to 64×64 pixels, see Fig. 1).

Results

Our experiment was composed of two parts: (1.) During **model learning**, we simulated daily human experience by training the model on 500k video frames of 40 speakers. (2.) For **model evaluation**, we emulated **experimental conditions** (Maguinness et al., 2021) and introduced 16 new speakers, with 8 speakers whose faces were never shown (black screen). We trained models and classifier for 100k more frames. Training was interspersed with **auditory-only test phases** to analyze speaker identification given voice only.

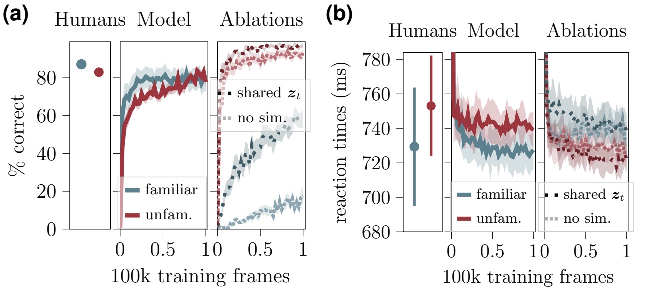

Figure 3: **Voice-based identification of speakers** with familiar or unfamiliar faces. Results from similar experiments in humans (Maguinness et al., 2021) are compared to our model and ablations, i.e., without simulation or with shared codes $\boldsymbol{z}_t$ for both modalities (Wu et al., 2023). We plot % correctly identified speakers **(a)** and reaction times **(b)** (± standard error).

Fig. 2 shows the **reconstruction of simulated faces** during testing, when *only perceiving auditory speech*. For speakers that were trained with the face (left), the model quickly ($t \approx 2$) predicted the correct face purely from the voice. For unfamiliar faces (right), the model simulated random, sex-specific facial features or the faces of similarly sounding speakers.

Face simulations strongly affected speaker identification. Fig. 3 shows the % correctly identified speakers and reaction times during testing (auditory-only) over training. As for humans in a similar experiment (Maguinness et al., 2021), our model learned to identify speakers with familiar faces (blue) with high accuracy and short reaction times. For unseen speakers (red), a similar accuracy required much more training. When cross-modal simulation was deactivated (no sim.) or observations from separate modalities were embedded into a shared code $\boldsymbol{z}_t$, as in the Rssm of Wu et al. (2023), the response pattern was reversed. These models learned to rely on vision only to identify familiar speakers and failed when this information was missing.

Conclusion

We have presented a first neuro-computational predictive model to explain face-benefits in human communication. Modality-specific codes, sensory substitutions, and internal face simulations were crucial to model the benefit.

References

Bhat, N. M. (2020). *pyttsx3 v.288.* Retrieved from `https://pypi.org/project/pyttsx3/` (used together with espeak v1.48.15 available over `https://espeak.sourceforge.net/index.html`)

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127-138. doi: 10.1038/nrn2787

Friston, K., Moran, R. J., Nagai, Y., Taniguchi, T., Gomi, H., & Tenenbaum, J. (2021). World model learning and inference. *Neural Networks*.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.*, *30*, 535–574.

Ha, D., & Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*.

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., & Davidson, J. (2019). Learning latent dynamics for planning from pixels. In *International conference on machine learning.*

Hafner, D., Lillicrap, T. P., Norouzi, M., & Ba, J. (2020). Mastering atari with discrete world models. In *International conference on learning representations.*

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Huang, Y., & Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580–593.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, *47*, 1122–1135.

Maguinness, C., Schall, S., & Kriegstein, K. (2021, 02). Prior audio-visual learning facilitates auditory-only speech and voice-identity recognition in noisy listening conditions. doi: 10.31234/osf.io/gc4xa

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169–181.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873–922.

von Kriegstein, K., Dogan, Ö., Grüter, M., Giraud, A.-L., Kell, C. A., Grüter, T., . . . Kiebel, S. J. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences*, *105*(18), 6747–6752.

Wu, P., Escontrela, A., Hafner, D., Abbeel, P., & Goldberg, K. (2023). Daydreamer: World models for physical robot learning. In *Conference on robot learning* (pp. 2226–2240).

Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., . . . Wang, F. (2023). Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8652–8661).