Emergence of in-context structure detection through self-supervised learning

Pierre Orhan (pierre.orhan@ens.psl.eu)
Laboratoire des systèmes perceptifs, 29 rue d'Ulm
Paris,France

Fosca Al Roumi (fosca.alroumi@cea.fr)
Cognitive Neuroimaging Unit, Neurospin Center
Gig/Yvette, France

Yves Boubenec (yves.boubenec@ens.fr)
Laboratoire des systèmes perceptifs, 29 rue d'Ulm
Paris,France

Jean-Remi King (jeanremi@meta.com)
Laboratoire des systèmes perceptifs, 29 rue d'Ulm
Paris,France

Abstract

Humans' ability to spontaneously detect symbolic structures is often considered to be essential to the acquisition of language and music. Prominent theories postulate that core, innate and *internal* mechanisms, like "merge" (Chomsky) or "neural recursion" (Dehaene), are foundational to this feat. Here we tested the alternative hypothesis that the ability to detect symbolic structures emerges from generic statistical learning operating onto *external* naturalistic inputs, that are structured in themselves. We focused on auditory stimuli, for which a wealth of experimental protocol questions structure detection. First, we exposed a self-supervised auditory model to a dataset merging music, speech and environmental sounds. Second, we exposed them to classical neuroscience experimental protocols and evaluated the models' ability to perform zero-shot detection of regularities, including algebraic structures. Like humans, training brought models to detect (1) repeated sequences, (2) probabilistic chunks and (3) algebraic structures, (4) with diminished performance for structures of increasing complexities. Furthermore, we show that this ability was a direct consequence of self-supervised learning: the more the models are exposed to natural sounds, the more they spontaneously detect increasingly complex structures. **Overall, we demonstrate that the emergence of the structure detection need not require a dedicated internal mechanism: rather, self-supervised learning operating on *external* sensory inputs being sufficient for the emergence of *internal* computations capable of detecting regular patterns such as algebraic structures.**

Keywords: Self-supervised learning — Statistical learning — Language — Music — Structure

Approach

The neural process underlying the human ability to detect symbolic structures, such as syntax in speech and geometrical symmetries in drawings remain unknown. Consequently, a variety of minimalist protocols have been proposed to test these views by isolating the brain and behavioral bases of structure building. To recapitulate these findings, we propose a novel approach, emphasizing models that learn from naturalistic stimuli and test their ability to detect artificial structures over a wide range of classical experimental protocols from the literature.

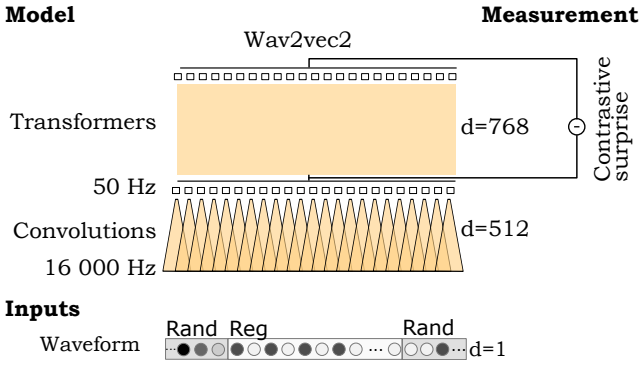

Figure 1: Diagrams of the model, measurement of the contrastive loss and structure of inputs.

Specifically, we pretrained a series of Wav2vec2.0 (Baevski et al. (2020)) deep neural models with self-supervised learning, to implicitly learn the latent structure of natural sounds. The pretraining dataset combines three different datasets: speech (Librispeech), music (FMA), and environmental sounds (Audioset from which we removed musical and speech sounds). Models are pretrained for 100 000 steps with base parameters (Fig.1). We then tested the model ability to perform structure detection at different training time on four experimental paradigms (Fig.2 A,D,G).

The experimental paradigms were all merged in a common framework, where a random sequence (Rand) was followed by a structured regular sequence (Reg), followed by another random sequences. For each sound, we measured the model surprise to each sound elements (tones or syllables). This

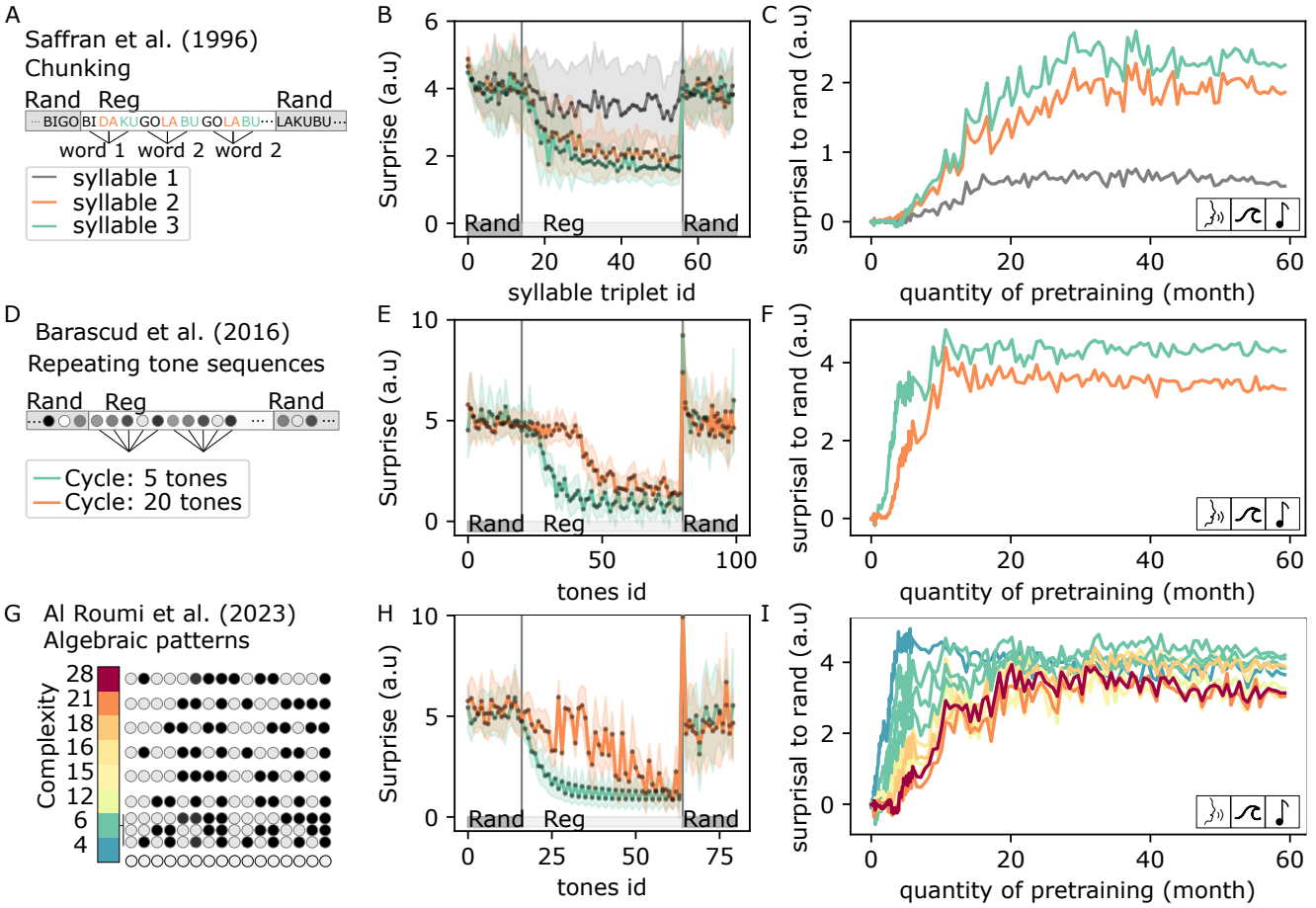

Figure 2: Testing structure detection and its emergence in three experimental protocol. First column: the three experimental paradigms. A: Words formed by three syllable randomly alternate between themselves, generating a regular stream of syllable. D: repetition of complex tones sequences. G: repetition of binary algrebraic tones sequences. Second column, B,E,H: contrastive loss of the model for each sound element in the stream. The contrastive loss decreases as soon as the repeated elements appear a second time in the regularity. The repetition period is termed cycle. Third column, C,F,I: difference of the contrastive loss on the Rand versus Reg stream as a function of the network pretraining.

surprise was the contrastive loss, computed by masking 20 ms (50Hz) latent vectors whose receptive field overlap with the sound element. If the model detected the sound structure, its surprise should progressively decrease during the regular sequence and suddenly peak at the onset of the second random sequence. To study the emergence of structure detection, the analysis was replicated on checkpoints logarithmically spanning the models pretraining.

Results We first investigated the models' ability to chunk sounds, first in speech and then sequences of tones. For speech, we replicated the experiment of Saffran et al. (1996), which demonstrated that 8-month-old children rapidly and spontaneously detect 3-syllable words in a stream of syllables. The stimulus switched from a random stream of syllables to a regular stream composed of successive 3-syllable words (Fig.2 A). During the regular stream, the model contrastive loss dropped progressively on the second and third syllables of each words (Fig.2 B). This demonstrates that the model was performing in-context spontaneous discovery of words. Remarkably, this ability emerged progressively during pretraining (Fig.2 C). Structure detection was also found for complex tone sequences investigated by Barascud et al. (2016) (Fig.2 D). Indeed the model contrastive loss dropped as soon as a series of 50-ms tones was repeated (Fig.2 E). The dynamics of the loss mirrored the human ability to optimally detect the embedded structure, as found by Barascud et al. Remarkably, structure detection emerged during pretraining (Fig.2 F). Finally, we tested if the model uses the sequence algebraic patterns to facilitate their detection. If this was the case, sounds with simpler structures should be discovered faster and more easily than sounds with complex structures. We thus studied binary tones sequences of increasing Language of Thought (LOT) complexity as introduced by Al Roumi et al. (2023) (Fig.2 G). The model was able to discover all structures (Fig.2 H), but this ability emerged earlier for simple sequences during pretraining (Fig.2 I).

The results demonstrate that structure detection is an

emerging property of a self-supervised model, resulting from the confrontation of generic-purpose statistical learning with the underlying structure of the outside world. Such models thus constitute a unifying computational framework for studying the emergence of structure discovery in neural networks.

References

Al Roumi, F., Planton, S., Wang, L., & Dehaene, S. (2023, November). Brain-imaging evidence for compression of binary sound sequences in human memory. *eLife*, *12*, e84376. Retrieved from https://doi.org/10.7554/eLife.84376 (Publisher: eLife Sciences Publications, Ltd) doi: 10.7554/eLife.84376

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020, October). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.* arXiv. Retrieved 2022-07-07, from http://arxiv.org/abs/2006.11477 (arXiv:2006.11477 [cs, eess]) doi: 10.48550/arXiv.2006.11477

Barascud, N., Pearce, M. T., Griffiths, T. D., Friston, K. J., & Chait, M. (2016, February). Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. *Proceedings of the National Academy of Sciences*, *113*(5), E616–E625. Retrieved 2023-09-04, from https://www.pnas.org/doi/10.1073/pnas.1508523113 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1508523113

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996, December). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. Retrieved 2023-07-20, from https://www.science.org/doi/abs/10.1126/science.274.5294.1926 (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.274.5294.1926