# A window on human and artificial cognition with reverse correlation

Etienne Thoret

https://doi.org/10.1038/s44159-023-00239-z

# A window on human and artificial cognition with reverse correlation

Etienne Thoret[1,2,3,4]

[1]Perception, Representations, Image, Sound, Music, UMR7061, CNRS, Marseille, France.
[2]Laboratoire d'Informatique et Systèmes, UMR7020, CNRS, Marseille, France.
[3]Institut de Neurosciences de la Timone, UMR7289, CNRS, Marseille, France.
[4]Institute of Language Communication and the Brain, Aix-Marseille University, Marseille, France.

Humans are experts in recognition tasks such as recognizing the emotion shown in a face. A central question in psychology is what sensory information humans use to achieve these feats. For instance, researchers might wonder which part of a face allows recognition of the emotions of conspecifics or what type of prosody characterizes trustworthy speech. These are the kind of questions about recognition that cognitive scientists have addressed for hundreds of years.

Dozens of methods have been invented to investigate visual or auditory recognition. Most often, scientists start by hypothesizing about specific features that are relevant. Then, they generate stimuli with this reduced set of varying features and test how these stimuli influence participants' recognition. Although efficient, this reductionist and hypothesis-driven approach limits the generalizability of conclusions because only a small number of features or factors can be tested. Along with many other researchers, I have extensively used this approach to investigate perception and cognition. However, I have always been frustrated by these limitations.

During my first post-doctoral position, I discovered that Frédéric Gosselin and Philippe Schyns had invented a method that circumvents these flaws. In their method, only a small randomly chosen part of the stimulus is made visible to the participant in the recognition task. They demonstrated the method in an emotion recognition task. They occluded the entire face except for small circular areas that they called 'bubbles'. In each trial, the positions of the bubbles were randomly shuffled. By running many trials and then correlating the correct and incorrect recognition responses with the positions of the bubbles, they revealed the parts of the faces that are most informative for correct recognition. For instance, they revealed that the information that drives the recognition of happiness is localized around the mouth. Contrary to most experiments, the bubbles technique does not require any hypotheses about the location of the relevant visual information.

This method was an immediate game changer in the study of both human and animal vision — even with pigeons. Variations of the method that use other ways of randomly perturbating the stimuli have been developed and are now known as reverse correlation ('revcor'). Reverse correlation refers to the systems identification field that offers a formal mathematical context to these methods.

In the auditory domain, this same method has been successfully applied to speech. As for images, bubbles can be used to randomly reveal otherwise-occluded acoustic information of auditory stimuli. In practice, the random perturbations are applied to signal processing representations such as the sound's spectrograms or speech synthesis models. For instance, it has been used to reveal the spectral features that characterize a 'smiling' voice. I had the opportunity to investigate the acoustic features that drive the identification of musical instrument timbre with this method.

**Corresponding author:** etienne.thoret@univ-amu.fr

https://doi.org/10.1038/s44159-023-00239-z

This versatile data-driven approach has had a striking impact on the vision and hearing sciences and has profoundly shaped my personal research. I am now applying this method to the field of explainable artificial intelligence. Artificial agents such as deep neural networks have raised numerous ethical questions regarding their inherent biases and the level of trust that one can place in them to replace human expertise. Because deep neural networks are considered to be models of human cognitive and neural processes, probing them with techniques such as bubbles and reverse correlation — which were initially designed for probing human performance — should help to quantify their similarity to humans.

**Competing interests:** The author declares no competing interests.

**Original article:** Gosselin, F. & Schyns, P. G. Bubbles: a technique to reveal the use of information in recognition tasks. Vision Res. 41, 2261–2271 (2001)

**Related articles:** Ponsot, E., Arias, P. & Aucouturier, J. J. Uncovering mental representations of smiled speech using reverse correlation. J. Acoust. Soc. Am. 143, EL19–EL24 (2018); Thoret, E. et al. Probing machine-learning classifiers using noise, bubbles, and reverse correlation. J. Neurosci. Meth. 362, 109297 (2021)

**Acknowledgments:** Research supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

**Corresponding author:** etienne.thoret@univ-amu.fr