Correspondence

The challenge of decoding animal communication using AI

Mor Taub[1,8], Inbal Arnon[2,3,8], Amiyaal Ilany[1,4,5,8], Mirjam Knörnschild[6,7,8], Yoav Ram[1,5,8], and Yossi Yovel[1,5,8,*]

Understanding animal communication requires identifying how signals map onto meaning in the receiver's perceptual space, yet most studies rely on clustering signals by physical similarity[1]. It is tempting to assume that the physical and the perceptual spaces are highly correlated, but this may not be the case. We exemplify this by examining a form of communication for which we *do* have some access to the perceptual space — the communication of pre-linguistic human toddlers directed towards adult humans. We use a small data set of vocalizations recorded in three different semantic contexts which could be described as: (1) distress calls; (2) contact calls with a specific recipient (father or mother); or (3) begging calls (emitted in expectation of feeding). Our results show that acoustic similarity does not necessarily correlate with the receiver's perceptual space.

A visual inspection of a subset of these vocalizations (Figure 1A,B) demonstrates how difficult it would be to group them properly based on spectrogram visual similarity. On the other hand, listening to the vocalizations, with our human brain (Audio S1–S3 in the Supplemental Information), suggests that the vocalizations in each column are all members of the same human perceptual context. This is because the human brain, which is the evolutionary target-receiver of these vocalizations, is highly tuned to extracting very specific features from the acoustic signal. We applied three models to embed the vocalizations in feature space, i.e., the model's equivalent of a perceptual space: a classical spectrogram signal processing approach previously applied to animal vocal communication[2]; a deep neural network (AVES) trained with self-supervision (without context annotations) on animal vocalizations; and a deep neural network (WavLM) trained with self-supervision on adult human speech. For comparison, we used a set of adult human speech of similar duration.

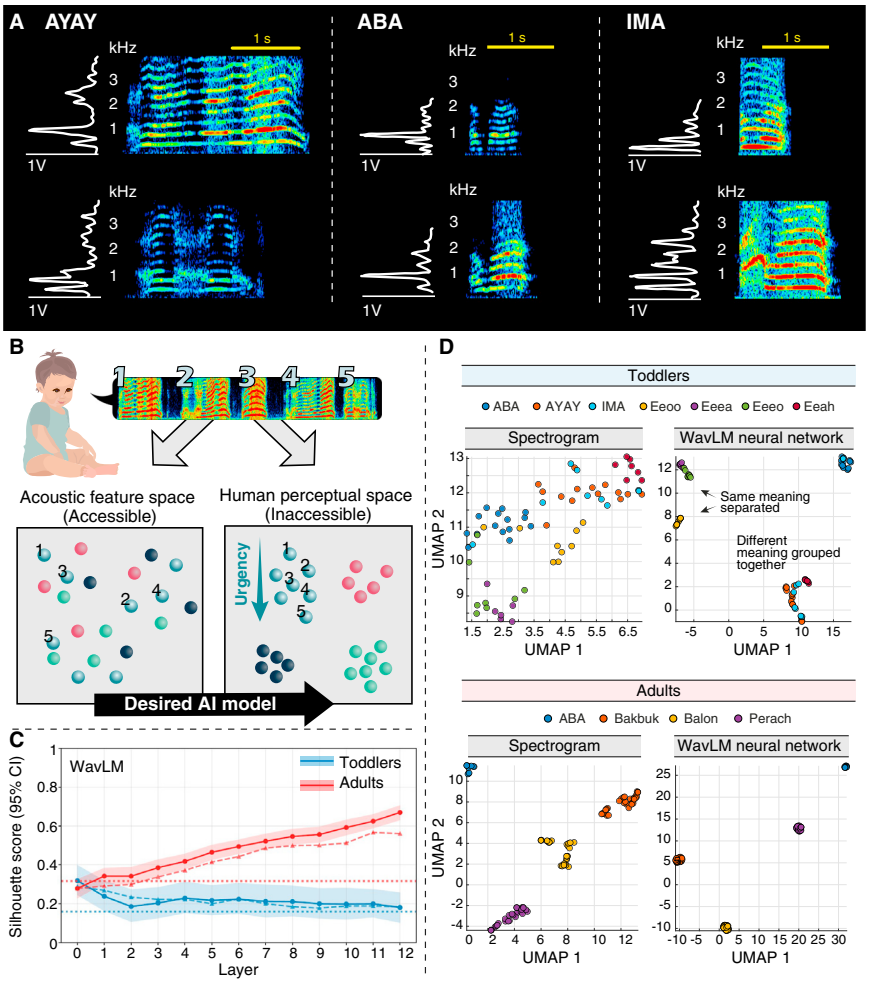

Figure 1. The difference between physical and perceptual spaces.
(A) Spectrograms and spectra of three sequences of syllables uttered by an 18-month-old toddler. Each column shows two spectrograms of syllables whose phonetic transcription appears above (AYAY, ABA, IMA). The two right columns call for a specific recipient: 'Dad' and 'Mom', respectively. The syllable on the left, on the other hand, is novel and toddler-specific. (B) A schematic of feature space versus perceptual space. A toddler emits a sequence of vocalizations (light blue). Observing these vocalizations in the accessible acoustic feature space (left) does not reveal the relevant relations between them. Only in the human perceptual space (right) do these vocalizations cluster properly as well as all other contexts (other colors). In perceptual space all blue vocalizations belong to one cluster and are ordered along a certain axis. We present the clusters as discrete, but in reality they might form a continuum. (C) Silhouette scores (for WavLM) measure how well samples (i.e., vocalizations) are assigned to their contextual cluster ('1' is best). The plots demonstrate that adult speech is clustered much better than toddler vocalizations, and that the gap increases for advanced model layers, which represent the perceptual space (early ones represent the acoustics). The results of a spectrogram-based model are depicted by a dashed horizontal line — and are already better for adult speech. Shaded areas represent 95% CI from (non-parametric) bootstrap. (D) UMAP showing the two-dimensional embedding of the best separating layer of two models.

While the classical spectrogram approach failed to represent the toddler vocalizations in feature space in a way that captures meaning, the deep neural networks performed much better, but still exhibited both types of error, grouping together vocalizations with different meaning and separating vocalizations with similar meaning

(Figure 1C,D). This was the case even for our ultra-simple dataset with minimal variance, while in reality, such vocalizations can be individual-specific and might sound differently under the exact same context when uttered by a different toddler or even by the same toddler on a different day. Such variance is also characteristic of many animal communication systems where much inter- and intra-individual variability can stem from various factors including sex, rank, region, health and idiosyncratic individual-unique signals[3].

All models performed much better on adult human speech and, as expected, the model trained specifically on human speech performed best (Figure 1C,D). Notably, the fact that even models that were not trained on human speech perfectly separated adult words suggests that speech has evolved features that are separable already in the acoustic space. The sequences of toddler vocalizations also provide insight into the challenges of discovering how vocalizations are concatenated into sequences[4] to express meaning. Listening to these sequences conveys a sense of increased urgency. This sensation is a result of subtle modifications to the acoustics that our human brain is tuned to extract. None of the tested models ordered the vocalizations according to their order of emission. Only a model that projects the vocalizations into the correct perceptual semantic space would also map them in a way that correlates with such concatenation regularities, for example along an axis of urgency (Figure 1B).

Every species requires its own species-specific perceptual model suitable for the communication modality it relies on. For example, in frogs and fruit flies female preference of male courtship 'songs' has been shown to be non-linear and that a small shift in the acoustics could be translated to a large shift in preference[5,6]. In both cases, decoding the meaning of communication would have been largely missed if only the acoustics, without the behavioral response, had been considered.

How can we approximate an animal's perceptual space? As argued by Nagel, Wittgenstein and others, we will never gain full access to the perceptual space of another individual let alone another species. However, there are ways to get closer to approximating an animal's perceptual space. Careful annotation of behavioral context by experts can yield insight about the animal's perceptual space. Two examples include an annotated set of fruit bat vocal interactions based on visual contexts[7] and of chimpanzee gestures[8]. Naturally, such efforts will always suffer from human biases.

Experiments that quantify the animal's response to measurable changes in communication signals, via playback or under natural conditions, are also crucial[9], but behavioral responses are not always explicit. Neural recordings or brain imaging while the animal is exposed to communication signals have revealed that specialized auditory neurons in the female fruit fly brain preferentially respond to species-specific patterns[10]. Notably, even a full recording of all neurons in the brain might not provide a sufficient representation of the perceptual space because it is not known how neural activity is integrated to generate perception.

Moreover, to gain a full understanding of a communication system, in addition to decoding the meaning of communication, models would need to also identify the structure of communication, that is, to identify the relevant units of the system and the regularities regarding how to concatenate them[4]. To further complicate things, the perceptual space may change based on state and context.

In summary, by using human toddler communication which we have a partial capacity to decode, we have found that clustering of non-human communication signals based on their physical properties may not correlate with the animal's perceptual space, the only relevant space to infer meaning. In this sense, communication can be viewed as one domain within an animal's '*Umwelt*' — the species-specific perceptual world through which signals become detectable, salient, and meaningful.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ACKNOWLEDGMENTS

We would like to thank Yosef Prat for assistance with adult vocalization acquisition.

SUPPLEMENTAL INFORMATION

Supplemental information including experimental procedures, data and code availability, and three audio files can be found with this article online at https://doi.org/10.1016/j.cub.2026.06.064.

REFERENCES

1. Zandberg, L., Morfi, V., George, J.M., Clayton, D.F., Stowell, D., and Lachlan, R.F. (2024). Bird song comparison using deep learning trained from avian perceptual judgments. PLoS Comput. Biol. *20*, e1012329.
2. Sainburg, T., Thielk, M., and Gentner, T.Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. PLoS Comput. Biol. *16*, e1008228.
3. Arnon, I., Kirby, S., Allen, J.A., Garrigue, C., Carroll, E.L., and Garland, E.C. (2025). Whale song shows language-like statistical structure. Science *387*, 649–653.
4. Suzuki, R., Buck, J.R., and Tyack, P.L. (2006). Information entropy of humpback whale songs. J. Acoust. Soc. Am. *119*, 1849–1866.
5. Baugh, A.T., Akre, K.L., and Ryan, M.J. (2008). Categorical perception of a natural, multivariate signal: Mating call recognition in túngara frogs. Proc. Natl. Acad. Sci. USA *105*, 8985–8988.
6. Talyn, B.C., and Dowse, H.B. (2003). The role of courtship song in sexual selection and species recognition by female Drosophila melanogaster. Anim. Behav. *68*, 1165–1180.
7. Prat, Y., Taub, M., and Yovel, Y. (2016). Everyday bat vocalizations contain information about emitter, addressee, context, and behavior. Sci. Rep. *6*, 39419.
8. Hobaiter, C., and Byrne, R.W. (2014). The meanings of chimpanzee gestures. Curr. Biol. *24*, 1596–1600.
9. Elie, J.E., de Witasse-Thézy, A., Thomas, L., Malit, B., and Theunissen, F.E. (2025). Categorical and semantic perception of the meaning of call types in zebra finches. Science *389*, 1210–1215.
10. Deutsch, D., Clemens, J., Thiberge, S.Y., Guan, G., and Murthy, M. (2019). Shared song detector neurons in Drosophila male and female brains drive sex-specific behaviors. Curr. Biol. *29*, 3200–3215.e5.

[1]School of Zoology, Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel. [2]Psychology Department, Hebrew University, Jerusalem, Israel. [3]School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK. [4]The Steinhardt Museum of Natural History, Tel Aviv University, Tel Aviv, Israel. [5]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel. [6]Museum für Naturkunde, Leibniz-Institute for Evolution and Biodiversity Science, Berlin, Germany. [7]Evolutionary Ethology, Institute for Biology, Humboldt-Universität zu Berlin, Berlin, Germany. [8]All authors contributed equally to the study.
*E-mail: yossiyovel@gmail.com