Optimal stimulus selection for dissociating acoustic and semantic processing of natural sounds

Maria Araújo Vitória (maria.araujo.vitoria@maastrichtuniversity.nl)
Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University
Maastricht, The Netherlands

Marie Plegat (marie.plegat@univ-amu.fr)
Institut de Neurosciences des Systèmes, UMR 1106, Inserm and Université Aix-Marseille, Marseille, France

Giorgio Marinato (giorgio.marinato@univ-amu.fr)
Institut des Neurosciences de La Timone, UMR 7289, CNRS and Université Aix-Marseille, Marseille, France

Michele Esposito (m.esposito@maastrichtuniversity.nl)
Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University
Maastricht, The Netherlands

Christian Herff (c.herff@maastrichtuniversity.nl)
Department of Neurosurgery, School for Mental Health and Neuroscience, Maastricht University
Maastricht, The Netherlands

Bruno L. Giordano (bruno.giordano@univ-amu.fr)
Institut des Neurosciences de La Timone, UMR 7289, CNRS and Université Aix-Marseille, Marseille, France

Elia Formisano (e.formisano@maastrichtuniversity.nl)
Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University
Maastricht, The Netherlands

Abstract:

Computational model-based analyses of behavioral and neural responses to natural sounds offer insights into the acoustic-to-semantic transformations involved in sound recognition. However, the inherent relation between low-level/intermediate features and semantic dimensions in natural stimuli complicates interpretation. Here, we present a method to identify optimal sets of natural sounds, minimizing the dependence between modeled representations at acoustic/intermediate and semantic level. We applied this approach in a behavioral experiment where participants made pairwise similarity judgments of sounds and semantic labels describing them. Our findings demonstrate that sound similarity judgements were most accurately modeled by the intermediate layers of a sound-to-event DNN (Yamnet), with minimal contribution of a semantic model (word2vec), whereas semantic similarity judgements exhibited the opposite pattern. These results highlight the effectiveness of our approach in dissociating acoustic and semantic processing of natural sounds, providing a framework for investigating further the neural computations underlying the processing of such stimuli.

Keywords: auditory semantics; AI-based modeling, real-world neuroscience, behavior

Introduction

Deriving meaning from sounds is crucial for comprehending our environment, yet the mechanisms by which acoustic information is transformed into meaningful semantic representations remain incompletely understood. Recent studies have utilized deep neural networks (DNNs) trained on auditory tasks to model behavioral and brain (fMRI) responses to natural sounds (Giordano et al., 2023; Kell et al., 2018; Tuckute et al., 2023). Giordano et al. (2023) compared the contribution of acoustic, semantic (NLP) and sound-to-event DNNs representations in modeling auditory perception. Among their findings, it was shown that sound dissimilarity judgements were predominantly predicted by sound-to-event DNNs, especially within intermediate layers of Yamnet (Gemmeke et al., 2017), https://github.com/tensorflow/models/tree/master/research/audioset/yamnet), while semantic models

(word2vec, Mikolov et al., 2013) also accounted for a significant portion of variance. However, whether this latter observation truly reflects semantic-level sound processing or is influenced by the inherent association between intermediate (referred to herein as *hyperacoustic*) and semantic representations remains unclear.

In this study, we develop an approach to identify optimal sets of natural stimuli, minimizing the dependence between alternative model representations and thus the efficiency of the experimental design, e.g. within linear modeling statistical frameworks (Mumford et al., 2015). Specifically, here we aim to identify a set of 150 natural sounds, minimizing the (linear and non-linear) relation between *hyperacoustic* and *semantic* sound representations. We employ this optimized stimulus set in a behavioral paradigm in which participants rated the similarities between pairs of sounds and semantic labels. Results demonstrate the effectiveness of our approach in dissociating acoustic and semantic processing of natural sounds.

Methods

Initial stimulus set

An initial set of sounds was obtained from FSD50K, an open dataset containing 51,197 audio clips, human labeled across over 350 classes from the AudioSet Ontology (Fonseca et al., 2022, https://zenodo.org/record/4060432). The selection process involved listening to sounds to ensure they could be distinctly identified as belonging to one basic label (i.e "guitar playing"); labels indicating superordinate level categories (i.e. "music playing") were excluded. This selection procedure resulted in a final set of 1,377 sounds, organized in 240 categories, each with five to six exemplars. For each sound category, a standardized semantic label was established, consisting of one noun describing the sound source ("who/what" sound descriptor) and one verb describing the sound mechanism/action ("how" sound descriptor, Giordano et al., 2022).

Selected sounds were preprocessed to ensure a duration between one and two seconds, for maximum subjective identifiability, then normalized to achieve the same peak of the time varying root mean square (RMS).

Model based stimulus selection

We selected an optimal subset of 150 sounds from the initial set of 1377 through an analytical procedure. Initially, we computed vector representations for each sound and linguistic label using Yamnet's ReLu layer 8 (considering the first 0.96s of sounds) and Word2Vec, respectively. We selected these representations as indicative of *hyperacoustic* and *semantic* processing based on a prior study (Giordano et al., 2023). We then formulated the optimal stimulus selection as the problem of identifying a subset of 150 stimuli that minimizes the dependence between these two vector representations. As a metric of dependence between vector spaces, we employed the *distance covariance* (dCov, Szekely et al., 2008), calculated from pairwise cosine distances within the vector representations. Subsequently, we obtained the optimal stimulus set as the solution to a mixed binary quadratic minimization problem (MBQP), solved using CPLEX (IBM). Alongside dCov minimization, we introduced the additional constraint that no more than two of the selected stimuli belonged to the same category.

Behavioral experiment

A behavioral experiment was conducted (N=7 participants), which included two distinct tasks. In the *sound similarity task*, participants judged the similarity of paired sounds by moving a slider along a scale marked with "very different" and "very similar" at the two extremes. The same rating method was adopted by participants in the *semantic similarity task* to judge the similarity of paired semantic labels. The sound (label) pairs were randomly derived from the optimized subset of 150 sounds. Each participant was presented with 300 pairs of sounds and 300 pairs of semantic labels, the same pairs were used across both conditions. The order of the tasks was counterbalanced across participants. All participants had normal hearing.

Comparison between behavioral and computational models

To assess the relation between behavioral results and model representations, we computed Pearson correlations between perceived similarity judgements for sound or semantic tasks and cosine similarity of Yamnet and word2vec embeddings. Additionally, layer-by-layer correlations were performed for Yamnet embeddings.

Results

The dCov sum of squares for 150 sounds resulting from the MBQP solution was much smaller (6.91), compared to those random selections with the same constraints of no more than two selected stimuli from the same category (36.13 ± 4.72, n=8925)).

The analysis of behavioral data demonstrates a high correlation between sound similarity and Yamnet embeddings and a small correlation with Word2Vec embeddings (Figure 1 a). Conversely, semantic

similarity judgments exhibit a stronger correlation with Word2Vec embeddings than with Yamnet embeddings (Figure 1 b). The dominance of the hyperacoustic model and the absence of semantic contribution was evident in 6 out of 7 participants during the sound similarity task (Figure 1 c-i). This validates our hypothesis that minimizing the dependence between Yamnet and Word2Vec embeddings for sound selection would enhance the dissociation between hyperacoustic and semantic representations.

Layer-by-layer correlation between Yamnet and behavioral similarities is illustrated in Figure 2. Overall, all layers are more correlated with sound similarity than semantic similarity judgements. Results confirm that intermediate layers (highest correlation in Relu layer 8) are more correlated than early or late layers (Figure 2 a).

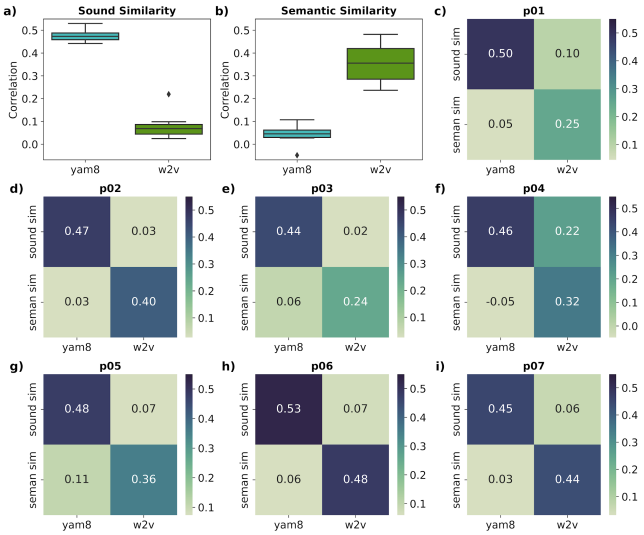

Figure 1. a-b) Correlation between sound similarity (a) or semantic similarity (b) and Yamnet (blue) and Word2Vec (green). c-i) Correlation between behavioral similarities models per subject.

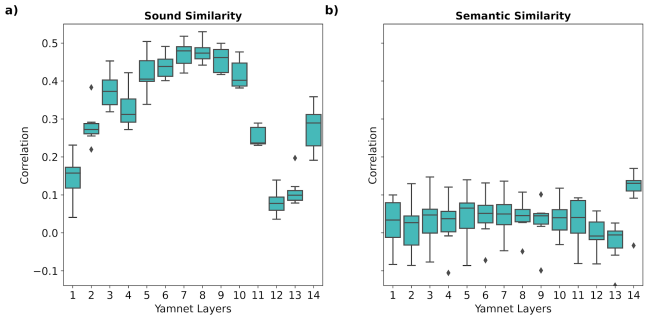

Figure 2. Correlation between behavioral sound similarity (a) or semantic similarity (b) and Yamnet layers' embeddings.

Conclusion

Our results demonstrate the efficacy of our analytical stimuli selection approach in disentangling acoustic and semantic representations. We are currently using the obtained stimulus set to collect brain responses (fMRI, MEG and intracranial EEG). and investigate the neural computations underlying the transformation of acoustic to semantic presentations.

Acknowledgments

This work was funded by the Dutch Research Council (NWO 406.20.GO.030 to EF), and by the French National Research Agency (ANR-21-CE37-0027-01, BLG; ANR-16-CONV-0002 ILCB; ANR11-LABX-0036 BLRI). The authors are thankful to Minne Pijfers for helping with the initial sound selection.

References

Fonseca, E., Favory, X., Pons, J., Font, F., & Serra, X. (2022). *FSD50K: An Open Dataset of Human-Labeled Sound Events* (arXiv:2010.00475). arXiv. http://arxiv.org/abs/2010.00475

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

Giordano, B. L., de Miranda Azevedo, R., Plasencia-Calaña, Y., Formisano, E., & Dumontier, M. (2022). What do we mean with sound semantics, exactly? A survey of taxonomies and ontologies of everyday sounds. *Frontiers in Psychology*, *13*. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.964209

Giordano, B. L., Esposito, M., Valente, G., & Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, 1–9. https://doi.org/10.1038/s41593-023-01285-9

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630-644.e16. https://doi.org/10.1016/j.neuron.2018.03.044

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

Mumford, J. A., Poline, J.-B., & Poldrack, R. A. (2015). Orthogonalization of Regressors in fMRI Models. *PLOS ONE*, *10*(4), e0126255. https://doi.org/10.1371/journal.pone.0126255

Szekely, G., Rizzo, M., & Bakirov, N. (2008). Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, *35*(6), 2769–2794. https://doi.org/10.1214/009053607000000505

Tuckute, G., Feather, J., Boebinger, D., & McDermott, J. H. (2023). Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biology*, *21*(12), e3002366. https://doi.org/10.1371/journal.pbio.3002366