Principled probing of foundation models in the auditory modality

Etienne Bost[1] **Mitsuko Aramaki**[2] **Richard Kronland-Martinet**[2]
Sølvi Ystad[2] **Thierry Artières**[1] **Thomas Schatz**[1]
[1]Aix Marseille Univ, Centrale Marseille, CNRS, LIS, Marseille, France
[2]Aix Marseille Univ, CNRS, PRISM, Marseille, France
{etienne.bost, thierry.artieres, thomas.schatz}@lis-lab.fr
{aramaki, kronland, ystad}@prism.cnrs.fr

Abstract

We leverage ecological theories of sound perception in humans [5, 6, 11] and a carefully designed dataset of perceptually calibrated sounds to develop and carry out principled fine-grained probing of foundation models in relation to the auditory modality. We show that internal activations of the state-of-the-art audio foundation model BEATs correlate better with perceptual dimensions than a supervised audio classification model and a text-audio multimodal model and that all models fail to represent at least one perceptual dimension. We also report preliminary evidence suggesting that directions aligning invariantly with a perceptual dimension can be identified within the representation space at inner layers of the BEATs model. We briefly discuss future work and potential applications.

1 Introduction

Recent attempts to replicate the success of text-based *foundation models* in other modalities or across multiple modalities have led to substantial advances in multiple domains, including computer vision and audio processing [17, 27, 2, 1]. Despite the unprecedented success of this approach, fine-grained probing of foundation models that take images as input (possibly among other modalities) reveals that they still suffer from many weaknesses and suggests possible avenues of improvement [14, 23, 22, 21]. Fine-grained probing of foundation models that take sounds as input (possibly among other modalities) have yet to be carried out to a similar extent, even though it is likely that it too would reveal substantial weaknesses and avenues of improvement.

Defining fine-grained, principled probing tasks in the auditory domain is challenging, however, and the evaluation of foundation models in this domain so far often revolves around limited tasks such as classification and captioning on poorly calibrated datasets [7, 16, 28]. In this study, we leverage theories of human sound source perception and a carefully designed dataset of perceptually calibrated sounds to develop and carry out principled fine-grained probing of foundation models in relation to the auditory modality. Specifically, we consider theories of ecological perception in the auditory modality, which posit the representation of a sound source along a set of perceptual dimensions, corresponding to perceived properties of the mechanical interaction producing the sound, such as the material of the interacting objects, whether they are empty or full, their spatial extension, etc. [5, 6, 11].

Beyond the identification of weaknesses and avenues of improvement in audio foundation models, our approach may help identify interesting dimensions within the usually opaque representation spaces of these models, with potential applications to model interpretability, conditional sound synthesis and computational modeling in cognitive neuroscience.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2 Related work

Probing of machine learning models has been conducted extensively across various domains, including natural language processing, computer vision, and speech processing, using both linear and non-linear classifiers [19, 26, 13], as well as discrimination metrics [25, 24, 30]. In the context of natural sounds—beyond speech and music—insights into the representations of foundation models have typically only been obtained through evaluation on standard classification or captioning benchmarks [7, 16, 28]. These evaluations, are coarse, however. They use roughly defined categories and do not leverage domain-specific knowledge. Recent studies have also explored the alignment of deep neural networks with human neural processing in the auditory modality, examining how well DNN audio models capture brain responses [29] and comparing invariances in artificial and biological networks [4]. To the best of our knowledge, however, we are the first to propose principled fine-grained probing tasks specifically designed for the study of natural sounds.

3 Materials and Methods

Probing sounds We use a perceptually-rated sound dataset collected at the PRISM laboratory[1], hereafter referred to as the *PRISM dataset*. It consists of 246 impact sounds recorded in an anechoic chamber and produced by striking various objects with a small hammer. Motivated by the ecological approach to sound perception [5, 6], the sounds are chosen to vary along three fundamental dimensions: the object's material (wood, glass, metal, stone, cardboard, ceramic or plastic), hollowness (whether it is full or not) and spatial extension (1D, 2D, or 3D). 17 human participants were recruited to classify each auditory stimulus based on perceived properties of hollowness, spatial extension, and material through auditory perception alone, without visual support. For each sound, we collected the distribution of participant categorical judgments across predefined categories for each perceptual property (material, hollowness, and spatial extension).

Tested models.

(i) **BEATs** is a foundation model trained exclusively on audio in a self-supervised manner on large-scale datasets [2]. Its architecture is transformer-based. BEATs is the current state-of-the-art for audio-only classification on the Audioset dataset, achieving a mean average precision (mAP) score of 0.506. Here, we leverage activations from BEATs' 12 self-attention layers to represent probing sounds.

(ii) **Yamnet** is an earlier convolutional neural network (CNN) model that achieved an mAP score of 0.314 on the Audioset classification task, on which it was trained in a supervised manner [9]. Yamnet provides a simple and reasonable baseline against which to compare BEATs. It consists of 14 depth-wise convolutional blocks, followed by a dense layer for classification. Here, we leverage activations from Yamnet's 14 layers to represent probing sounds.

(iii) **CLAP** (Contrastive Language-Audio Pretraining) is a foundation model linking audio with descriptive language through training on more than 5,000 hours of audio across 725,338 audio clips with matched text descriptions [3, 31]. Through contrastive learning, it establishes a shared multimodal space for audio and text, allowing it to perform well across 16 diverse tasks with impressive zero-shot capabilities, especially in sound event classification. CLAP architecture is transformer-based. Here, we leverage activations at the output of each of CLAPs' 12 successive transformer blocs to represent probing sounds.

4 Experiment 1: Comparing human and model similarity spaces

Approach The sound representations we can derive from the tested models and the human data available in the PRISM dataset are different. In models, we obtain vectors of neural activations at each layer in response to a given sound. In humans, we collect a distribution of category labels—one per participant—for each of three categorization tasks: material, hollowness, and spatial extension. In Experiment 1, we bridge this gap by computing dissimilarity matrices between sounds based either on model representations or human judgments and studying the correlation between the entries in these matrices [10]. We also add a third representation space based on the physical

[1] `https://www.prism.cnrs.fr`

ground truth (e.g. the actual material of a sound source, rather that the perceived material). For machines we compute a cosine distance, between a model's representations of each pair of probing sounds at each layer. For humans, for each pair of sound and each of the three categorization tasks, we compute either (i) a symmetric Kullback-Leibler (KL) divergence between the distributions of human responses for these two sounds in that task or (ii) a discrete distance between the modes of these distributions ($d = 0$ if the mode is the same, $d = 1$ otherwise). We also compute for both KL and discrete dissimilarities, an average over the three categorization tasks as an overall measure of similarity based on human judgments. Finally, we compute a physical similarity as a discrete distance between physical ground truth labels.

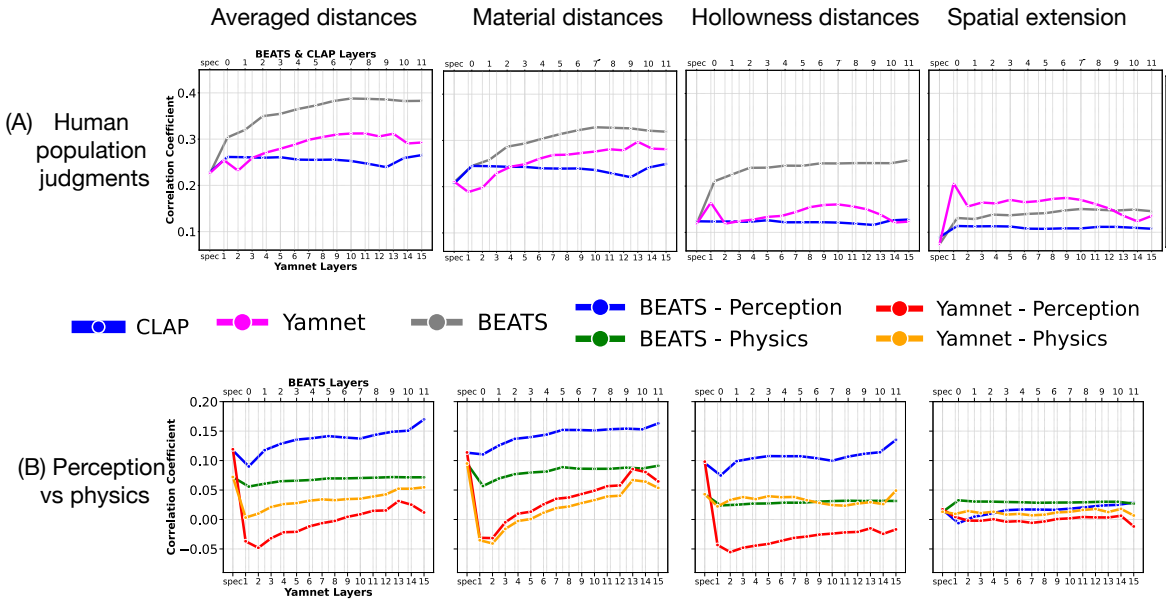

Figure 1: (A) Correlation between similarity in model representations and KL-divergence similarity in human judgments. (B) Correlation between similarity in BEATs and Yamnet representations and discrete distance similarity based either on human judgments or on physical ground truth. Note that for Yamnet, layer 15 is the output layer.

Results The spatial extension of objects (1D, 2D or 3D) appears to be overall poorly represented in all three model's layers (Figure 1A, last column). Hollowness is overall better represented in BEATs layers than in CLAP or Yamnet layers and, within BEATs, it is better represented in the deeper layers (Figure 1A, column 3). Material appears better represented in BEATs than in Yamnet and better represented in Yamnet than in CLAP, although the differences between the three are less marked than for the hollowness dimension. Within BEATs and Yamnet, but not within CLAP, material appears better represented in deeper layers (Figure 1A, column 2). Overall, the averaged similarity between human judgments on the three dimensions considered account for about $r^2 = .4^2 = 16\%$ of the variance in representational similarity in the upper layers of BEATs on the sounds from the PRISM dataset (Figure 1A, column 1). In the appendix (Figure A2), we show that the difference between BEATs and Yamnet on the material dimension is mostly due to a better preservation of spectral similarity information in BEATs inner layers. The superiority of the BEATs model on the hollowness dimension, however, is not explained away by regressing out spectral similarities, suggesting a more profound difference in what the two models learned with respect to that dimension.

In the remainder of the article we focus on the two most promising models, BEATs and Yamnet. As can be seen on Figure 1B, except for the spatial extension, which is poorly represented by all layers of all networks, inner layers of the BEATs model correlate substantially better with perceived than physical similarity. Remarkably, correlation with physical similarity tend to stay mostly flat as one goes towards inner layers, while correlation with perceived similarity increases. This suggests that BEATs may have learned a properly *perceptual* representation of sounds, which is all the more remarkable since it was only trained to predict upcoming sounds in a large database of natural

sounds, without access to any form of supervision or perceptual judgments. The results are much more contrasted for the Yamnet model—trained with explicit supervision on audioset—which only correlates slightly better with perceptual rather than physical similarity on the material dimension.

Discussion This first experiment suggests that the perceived spatial extension of sound sources is poorly represented in the tested models. These models might thus perform poorly in applications requiring fine timbre distinction related to this perceptual dimension.

Furthermore, the poor performance of the CLAP model compared to BEATs and Yamnet is surprising and interesting, especially since CLAP generally outperforms Yamnet, including on audio classification tasks. A possible interpretation may be that CLAP's difficulty in representing the fine-grained sound properties under study result from its multimodal training objective. Being trained for audio-text alignment, CLAP likely prioritize acoustic features that support this alignment over those intrinsic to audio alone. And because text contains less information than raw audio, this might create an "information bottleneck" that limit CLAP's capacity to encode detailed fine-grained audio structures. Our results thus suggest that although CLAP excels in broad classification and alignment tasks, its performance on generative or perceptual tasks that rely on subtle acoustic nuances, like the perceptual dimensions studied here, might be generally poor.

The best performing model is the purely audio self-supervised BEATs model. Our results are compatible with the presence of an explicit representation of perceived material and hollowness within the inner layers of the BEATs network. Correlation between similarity structure is necessary but not sufficient to support this idea, however, and does not give us actual access to that putative representation for use in practical applications. This motivates our next experiment, in which we test whether we can identify explicit axes within model representations that align with a perceptual dimension in a way that is invariant to changes along other perceptual dimensions.

5 Experiment 2: Finding perceptual dimensions in model representations

Approach In this section we want to test if we can find axes within BEATs representation space that provide an invariant index to perceptual dimensions. We provide preliminary results for the case of material. Our approach consists of training and testing linear classifiers of material (through $L2$-regularized multinomial logistic regression) on BEATs representations of a number of carefully selected subsets of the PRISM dataset. We are in particular interested in assessing whether a classifier trained on only sounds perceived to be of a specific type (e.g. only sounds perceived to be hollow or from 1d and 2d sources) *transfers* well to classifying the material of sounds of a different type. High transfer accuracy would suggest that the directions identified by the multiclass linear classifier index the perceptual dimension of material in a way that is invariant to changes along other perceptual dimensions (hollowness and spatial extension).

Results As can be seen on Figure 2, preliminary results suggest that it is possible to transfer material classification boundaries learned from sounds with specific characteristics to sounds with different properties. A proper statistical analysis of these results is warranted, but the accuracy scores consistently exceeding chance level and the moderate to inexistent drop in accuracy when compared to a baseline with no transfer (in purple) are promising. As in Experiment 1, a steady increase in performance is observed across the layers of BEATs, with deeper layers showing the highest accuracy. This aligns with the findings from the first experiment, where a stronger correlation between model representations and human perception of material was noted in BEATs' deeper layers.

Discussion Although preliminary, these results support the idea that it may be possible to find axes within BEATs representation space that provide an invariant index to fine-grained perceptual dimensions. This could have implications for practical applications where fine-grained distinctions matter—for example for sound effects indexation or synthesis. For example, one could imagine moving a sound along an "hollowness" direction to generate sounds that are more or less hollow. This could enable more intuitive and precise control of synthesis, without sacrificing the benefits of deep networks in synthesis, such as the sheer diversity of sounds that can be synthesized. Our next step will be to use general-purpose sound synthesis architectures, such as Audio-LDM [12] to test this idea.

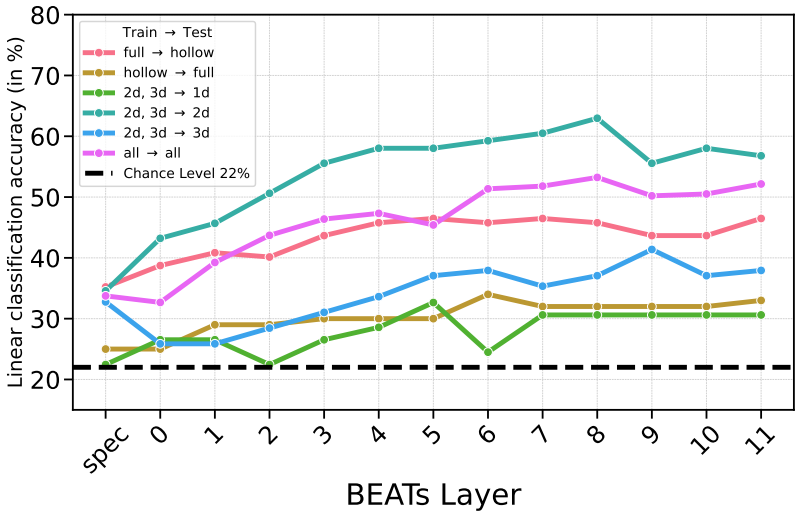

Figure 2: Test accuracy for multinomial logistic regression on the 8 material categories. The legend indicates the nature of the train/test split of the PRISM database (e.g. if 'Train' is 1d, 2d and 'Test' is 3d, it indicates that only sounds from objects in 1 and 2 dimensions were used for training and only sounds from objects of 3 dimensions were used for testing). Chance level corresponds to a classifier that always predicts the most common category, in this case metal with 22% of occurrences.

6 Conclusion

We showed that internal activations of the BEATs model correlate better with perceived material and hollowness than those of CLAP or Yamnet, that the multimodal text-audio CLAP model appears limited in representing fine-grained audio information and that all models fail to represent spatial extension. We also reported preliminary evidence suggesting that directions aligning with perceived material in a way invariant to variations along other perceptual dimensions can be identified within the representation space at inner layers of the BEATs model. Future work will investigate additional models and perceptual dimensions, including dimensions related not only to the attributes of sound sources, but also to the nature of the sound-producing interaction (e.g. impact, scraping, flowing, etc.).

Our approach may help identify model weaknesses and avenues for improvement. For the specific case of the failure to represent spatial extent, our next step will be to test models learning from multimodal input beyond text and audio like [1], which might perform better because spatial extension is much more salient, for example, in visual and haptic signals than in audio or text. Our approach may also be helpful for model transparency, by identifying interpretable structures within internal representations. Identifying directions corresponding to invariant perceptual dimensions in audio foundation models may also open up new avenues for designing intuitive sound indexation and synthesis tools. This may prove especially useful for applications where fine-grained audio distinctions matter. Indeed, our results with the CLAP model suggest that these applications may be difficult to handle with text-based prompting only. Finally, investigating which perceptual dimensions are represented by audio foundation models has bearing on issues of current interest in the computational cognitive neurosciences of auditory perception [18, 15, 8, 20].

Acknowledgments and Disclosure of Funding

This work received support from the French government under the France 2030 investment plan, as part of the Initiative d'Excellence d'Aix-Marseille Université – A*MIDEX AMX-21-PEP-021. This work, carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government (France 2030), managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

References

[1] Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., and Liu, J. (2024). Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36.

[2] Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., and Wei, F. (2022). Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

[3] Elizalde, B., Deshmukh, S., Al Ismail, M., and Wang, H. (2023). Clap: learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[4] Feather, J., Leclerc, G., Madry, A., and McDermott, J. H. (2023). Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11):2017–2034.

[5] Gaver, W. W. (1993a). How do we hear in the world? explorations in ecological acoustics. *Ecological psychology*, 5(4):285–313.

[6] Gaver, W. W. (1993b). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1):1–29.

[7] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

[8] Giordano, B. L., Esposito, M., Valente, G., and Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, 26(4):664–672.

[9] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

[10] Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249.

[11] Kronland-Martinet, Y. S. A. M., McAdams, R. S. K. S. C., and RR, S. P. A. F. (2019). Timbre from sound synthesis and high-level control perspectives timbre: Acoustics. *Perception, and Cognition*.

[12] Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y., and Plumbley, M. D. (2024). Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[13] Ngo, J. and Kim, Y. (2024). What do language models hear? probing for auditory representations in language models. *arXiv preprint arXiv:2402.16998*.

[14] Nikolaus, M., Salin, E., Ayache, S., Fourtassi, A., and Favre, B. (2022). Do vision-and-language transformers learn grounded predicate-noun dependencies? *arXiv preprint arXiv:2210.12079*.

[15] O'callaghan, C. (2011). Xiii—hearing properties, effects or parts? In *Proceedings of the Aristotelian Society (Hardback)*, volume 111, pages 375–405. Wiley Online Library.

[16] Oncescu, A.-M., Koepke, A., Henriques, J. F., Akata, Z., and Albanie, S. (2021). Audio retrieval with natural language queries. *arXiv preprint arXiv:2105.02192*.

[17] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

[18] O'Callaghan, C. (2021). Auditory Perception. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition.

[19] Raymondaud, Q., Rouvier, M., and Dufour, R. (2024). Probing the information encoded in neural-based acoustic models of automatic speech recognition systems. *arXiv preprint arXiv:2402.19443*.

[20] Robert, P., Zatorre, R., Gupta, A., Sein, J., Anton, J.-L., Belin, P., Thoret, E., and Morillon, B. (2024). Auditory hemispheric asymmetry for actions and objects. *Cerebral Cortex*, 34(7):bhae292.

[21] Salin, E. (2024). Etude de la compréhension multimodale des modèles transformeurs vision-langage.

[22] Salin, E., Ayache, S., and Favre, B. (2023). Towards an exhaustive evaluation of vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 339–352.

[23] Salin, E., Farah, B., Ayache, S., and Favre, B. (2022). Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[24] Schatz, T. (2016). *ABX-discriminability measures and applications*. PhD thesis, Université Paris 6 (UPMC).

[25] Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., and Dupoux, E. (2013). Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline. In *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, pages 1–5.

[26] Shah, J., Singla, Y. K., Chen, C., and Shah, R. R. (2021). What all do audio transformer models hear? probing acoustic representations for language delivery and its structure. *arXiv preprint arXiv:2101.00387*.

[27] Srivastava, S. and Sharma, G. (2024). Omnivec: Learning robust representations with cross modal sharing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1236–1248.

[28] Trowitzsch, I., Taghia, J., Kashef, Y., and Obermayer, K. (2019). The nigens general sound events database. *arXiv preprint arXiv:1902.08314*.

[29] Tuckute, G., Feather, J., Boebinger, D., and 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.

[30] Versteegh, M., Thiolliere, R., Schatz, T., Cao, X.-N., Anguera, X., Jansen, A., and Dupoux, E. (2015). The zero resource speech challenge 2015. In *Interspeech*, volume 15, pages 3169–3173.

[31] Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. (2023). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A Correlation between and within human judgments and physical labels

We verify here that human perception judgments on the different perceptual dimensions are not correlated with each other (Figure A1). We also measure the correlation between these judgments and the physical reality.

Human judgments do not appear to be strongly correlated with physical reality A1, highlighting the distinction between perception and physical properties. Specifically, there seems to be almost no correlation between judgments of hollowness and their physical counterparts. Interestingly, a notable correlation exists between the perception of hollowness and spatial dimension, which may be explained by the fact that hollow objects cannot exist in two dimensions. This relationship suggests a potential link between perceived hollowness and spatial extent that warrants further investigation.

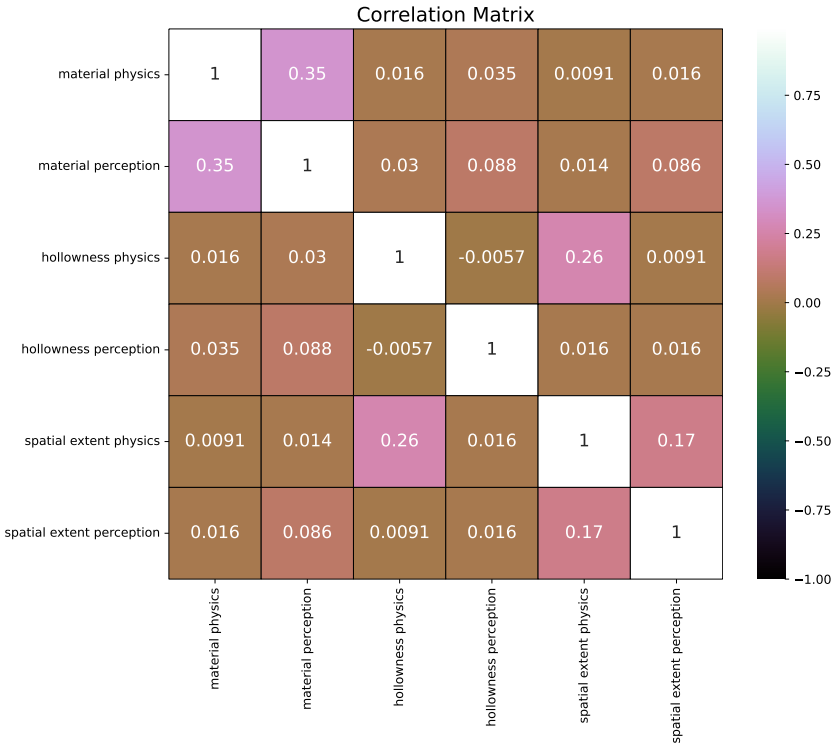

Figure A1: Correlation matrix for the different types of discrete distances (perceptual and physical) across the 3 perceptual dimensions

B Regressing out spectral similarity

In order to study the effects of the model beyond what can already be learned from low-level sound representations such as spectrograms, we sought to disentangle the two. We found that the distances between spectrograms are strongly correlated with the distances between representations in the intermediate layers of BEATs, but not with those of Yamnet (see Figure A2). This leads us to believe that a significant portion of the information present in the spectrogram is ignored from the earliest layers in Yamnet, whereas BEATs seems to better preserve useful information contained in the spectrogram. It is therefore reasonable to ask whether the lower performance of Yamnet in correlating with perceptual judgments could simply be explained by its disregard for spectral information regarding materials, dimensions, and hollowness, rather than by the model's ability to represent sounds accurately. To verify this, we replicate the previous experiments while removing the effect of the spectrogram on the representations by regressing it out. That is to say, we train a linear regression to predict the distances between representations for each layer of the models based on the distances between spectrograms, and we replicate the analysis from Figure 1 on the residuals from this regression.

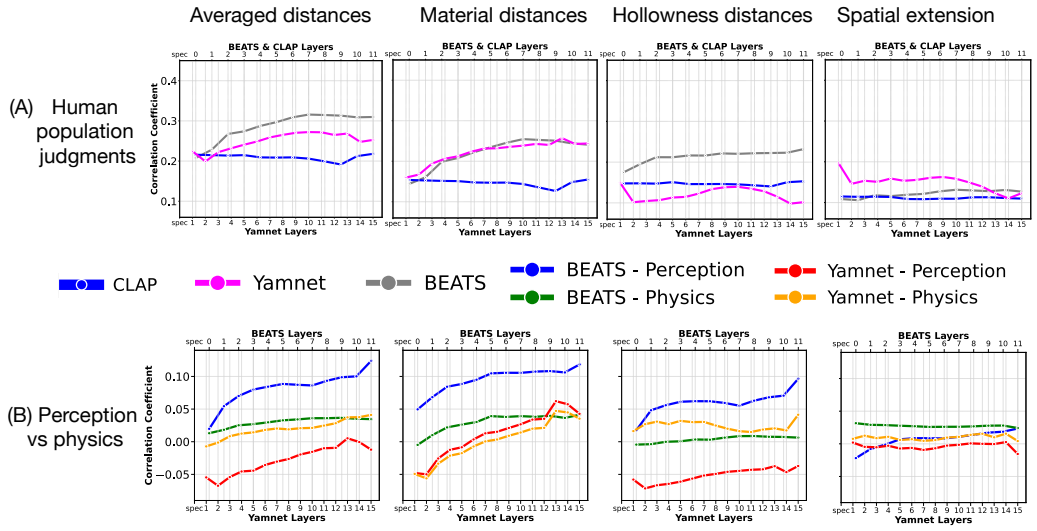

Figure A2: (A) Correlation between residual similarity in model representations after regressing out spectral similarities and KL-divergence similarity in human judgments. (B) Correlation between residual similarity in model representations after regressing out spectral similarities and discrete distance similarity based either on human judgments or on physical ground truth for the BEATs and Yamnet models.