Temporal dynamics of natural sounds representation in the human brain

Marie Plegat (marie.plegat@univ-amu.fr)
Institut des Neurosciences de La Timone, UMR 7289, CNRS and Université Aix-Marseille, Marseille, France.
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.

Christian Ferreyra (christian.ferreyra@etu.univ-amu.fr)
Institut des Neurosciences de La Timone, UMR 7289, CNRS and Université Aix-Marseille, Marseille, France.
Laboratoire d'Informatique et des Systèmes, UMR 7020, CNRS and Université Aix-Marseille, Marseille, France.

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

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

Daniele Schön (daniele.schon@univ-amu.fr)
Institut de Neurosciences des Systèmes, UMR 1106, INSERM 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.

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

Abstract

Acoustic and semantic representations involved in the temporal dynamics of the cerebral processing of natural sounds are often studied separately. As a consequence, we lack direct knowledge of how the human brain transforms complex acoustic waveforms into semantic representations of the acoustic environment. Here, we aimed to elucidate this process by predicting magnetoencephalographic (MEG) responses to natural sounds using acoustic, and semantic (text-based) models. Critically, we also consider two recently developed sound-processing convolutional neural networks (CNNs) that differ only in their loss function: CatDNN, which learns sound-event categories, and SemDNN, which learns continuous semantic embeddings (Word2Vec). We observe that DNNs better predict the dynamic MEG response, except at a long latency (800–1000 ms) where higher-level acoustics seems to dominate (auditory dimensions). Focusing on DNNs, we observe a potential switch from initial protoacoustic/categorical semantic representations (CatDNN, 250 ms) to more refined continuous semantic representations (SemDNN, 500–800 ms). Overall, our findings suggest limitations in the text-based modeling of the cerebral representations of natural sounds, and give a temporally resolved description of the cerebral dynamics of the acoustic-to-semantic transformation.

Keywords: Auditory processing ; Natural sounds ; Acoustic-to-semantic ; Artificial neural networks ; RSA

Introduction

Extracting meaning from sounds is essential for real-world behavior and relies on transforming the acoustic input to semantic representations. Recent studies have shown that event-classification convolutional neural networks (CNNs) better model auditory processing, but also highlight the need to disentangle acoustic and semantic components through careful model comparison. Giordano et al. (2023) used complex acoustic, semantic, and CNN models to clarify the contribution of each in modeling fMRI and behavioral responses to controlled natural sounds. They found predominance of acoustic over semantic models in Heschl's gyrus (HG), while both are equally well represented in the Superior Temporal Gyrus (STG). CNNs match acoustic model performance in HG, outperform all models in the acoustic-based dissimilarity task and in STG, but are outperformed by semantic models in the label-based dissimilarity task. A similarly controlled computational approach is still lacking, however, for detailing the temporal dynamics of cerebral representations (e.g., De Lucia et al., 2010; Lowe et al., 2023). To address this shortcoming, we carried out a study mapping acoustic, semantic, and CNN models of sound processing onto dynamic MEG responses to natural sounds.

Methods

Experimental design Each participant (N=21) listened to a common set of sounds (duration = 2 s each) eight times across two MEG sessions (total of 25,200 trials) while performing a one-back repetition detection task. The Common Set consisted in 150 sounds designed to minimize covariance between intermediate (eight layer of Yamnet network) and semantic (Word2Vec) models(Araújo et al., 2024). Each sound is labeled by its source (what/who), the action involved (how), and their combination (what/who-how) (Giordano, de Miranda Azevedo, Plasencia-Calaña, Formisano, & Dumontier, 2022). Additional sound sets, unique to different groups of participants, and presented throughout the same experiment are not considered for the results shown here.

Computational models We considered three classes of models: acoustic, semantic (text-based), and CNN models. Acoustic models approximating processing at various stages of the sound-processing hierarchy. The Cochleagram and Modulation Transfer Function approximated lower-level acoustic representations in subcortical structures and in the primary auditory cortex, respectively (Giordano, Esposito, Valente, & Formisano, 2023). The Auditory Dimensions model (AudDim), included multiple components that estimate psychoacoustic attributes (pitch, loudness, periodicity, brightness, and roughness) that better capture acoustic representations in the superior temporal gyrus (STG) (Giordano et al., 2023). Each of these models was estimated for three 1 s windows of the 2 s sound, starting at 0, 0.5, and 1 s from sound onset. Text-based semantic models: Word2Vec, and a co-occurrence model estimating semantic similarity based on the co-occurrence of text labels in a large database of descriptions of sound scenes. Both models included three separate components, focusing on what labels alone, how labels alone, or on the similarity between what and how labels. We finally considered two recently developed CNN models of sound processing (Esposito et al., 2024) differing only in the loss function: CatDNN, which learned sound-event categories (one-hot encoding), and SemDNN which learned continuous semantic embeddings (derived from Word2Vec). For both models, we estimated activations in each convolutional layer and the output considering three separate 1 s windows of the sound stimulus (windows beginning at 0, 0.5, and 1 s).

Model-representation analysis We considered a cross-validated representational similarity analysis framework (leave-one-participant-out) to predict the time-varying dissimilarity of sounds (CommSet) in MEG sensor space (MEG representational dissimilarity matrices – RDMs) considering the sound dissimilarities according to the computational models. MEG crossnobis distances (Walther et al., 2016) were estimated independently for each participant (leave-one-repetition-out with whitening based on MNE Python approach) by considering the 306 sensors as features (gradiometers and magnetometers). For all models except for the co-occurrence model, we considered as predictors both the Euclidean and cosine dissimilarity of sounds according to the features of each model component. For the co-occurrence model, we considered three separate approaches for deriving semantic

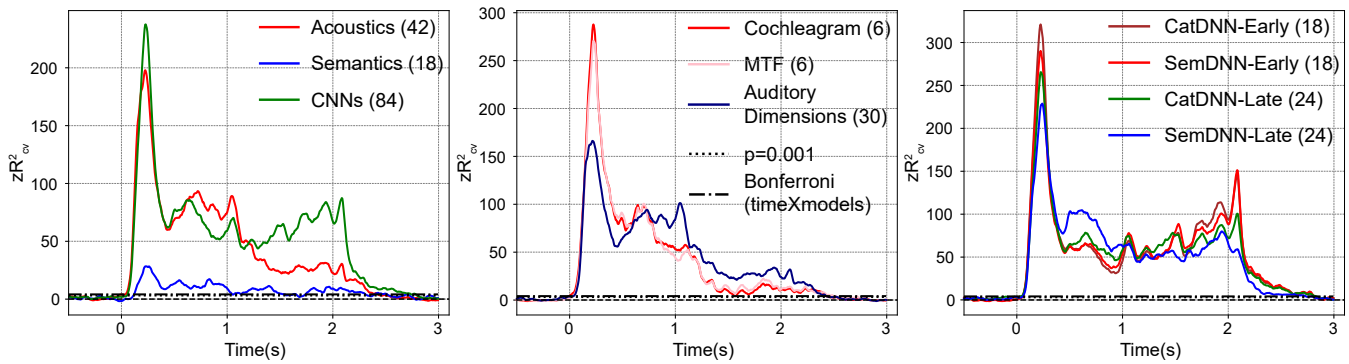

Figure 1: Computational representation of natural sounds in dynamic MEG responses. Sound duration = 2 s.

dissimilarities based on co-occurrences (negative of Jaccard coefficient, joint probability, and of pointwise mutual information).

Model representation analyses relied on a leave-one-participant-out approach for the prediction of MEG RDMs based on model RDMs (Giordano et al., 2023) within a regression framework. We followed the latest practice of whitening model and MEG RDMs to eliminate mathematical dependencies brought by the distance computation (e.g., the distance between sounds A and B is not independent of the distance between sounds A and C) (Diedrichsen et al., 2021). Finally, we implemented a novel Monte Carlo approach for eliminating biases in model-predictivity brought by differences in the number of predictors, resulting in model predictivity estimates standardized considering the mean and standard deviation of null-hypothesis Monte Carlo distributions (zR^2_{cv}).

Results

Figure 1 illustrates the predictive power of various classes of computational models. We include a provisional Monte Carlo significance threshold in both uncorrected form (p = 0.001), and Bonferroni-corrected across time-points and model classes within each panel. In the left panel, models are grouped into Acoustics, Semantics (text-based), and sound-to-semantic CNNs. Acoustic and CNN models outperform Semantic models (max zR^2_{cv}=25 for Semantics; mean zR^2_{cv}=50 for Acoustics, 75 for CNNs). The predictive power of both Acoustic and CNN models peaks at 250ms, but CNNs reach higher zR^2_{cv} values, suggesting that they do not merely capture the representation of acoustic features. From 800–1100ms, acoustic models better predict MEG dissimilarities than CNNs. After 1100ms, CNNs continue to perform well while the predictivity of Acoustics drops. In the middle panel, we detail the representation of the three acoustic models. The post-onset prediction peaks at 250ms and is mainly driven by the Cochleagram and MTF models (max zR^2_{cv}=275), while the AudDim model shows a weaker predictive power (max zR^2_{cv}=160). After this peak, Cochleagram and MTF predictions gradually decline, while AudDim shows a predictive advantage in the 800–1100ms window. A longer latency for the representation of the AudDims model than for the Cochleagram and MTF models is consistent with the representation of these acoustic attributes in the post-primary auditory cortex (STG), and highlights their potential higher-level nature and computation based on more finely grained representations such as the MTF. In the right panel, we contrasted the representation of the first two convolutional layers of each CNN (CatDNN-Early and SemDNN-Early) with that of the subsequent three layers, which are more strongly diversified between the two networks (results not shown). As expected, early CatDNN and SemDNN layers are characterized by similar peak levels of MEG predictivity, with a small advantage for CatDNN-Early predictions. More importantly, SemDNN-late outperforms all models between 500–800ms, suggesting a potential switch from early categorical (/acoustic) representation to later continuous semantic representations.

Conclusions

Our results show the ability of CNNs to predict MEG response dissimilarities by potentially capturing both acoustic and semantic underlying processes, while text-based models fail to reflect the brain's semantic processing of natural sounds. Importantly, text-based models of sound processing, including Word2Vec, capture less accurately MEG responses to natural sounds than CNNs that learn Word2Vec directly from sound (SemDNN). This suggests that the cerebral representation of the semantics of natural sounds overlaps only in part with that of linguistic semantics.

Acknowledgments

This work was funded by the French National Research Agency (ANR-21-CE37-0027-01, BLG; ANR-16-CONV-0002 ILCB; ANR11-LABX-0036 BLRI), by the Dutch Research Council (NWO 406.20.GO.030 to EF), and by ERC-2024-SyG NASCE (Proj. 101167313) to EF and BLG. The authors thank the Institut du Cerveau (CENIR, Paris, France) for enabling the MEG acquisitions.

References

Araújo, M., Plegat, M., Marinato, G., Esposito, M., Herff, C., Giordano, B. L., & Formisano, E. (2024). Optimal stimulus selection for dissociating acoustic and semantic processing of natural sounds. *Conference on Cognitive Computational Neuroscience*.

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature reviews. Neuroscience*, *14*.

Diedrichsen, J., Berlot, E., Mur, M., Schütt, H. H., Shahbazi, M., & Kriegeskorte, N. (2021). Comparing representational geometries using whitened unbiased-distance-matrix similarity. *eural Data Science*, *2*.

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology*, *13*.

Esposito, M., Valente, G., Plasencia-Calaña, Y., Dumontier, M., Giordano, B. L., & Formisano, E. (2024). Bridging auditory perception and natural language processing with semantically informed deep neural networks. *Scientific Reports*, *14*.

Giordano, B. L., de Miranda Azevedo, R., Plasencia-Calaña, Y., Formisano, E., & Dumontier, M. (2022). What do we mean with sound semantics, exactly? asurvey of taxonomies and ontologies of everyday sounds. *Frontiers in Psychology*, *13*.

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*, *26*.

Giordano, B. L., Whiting, C., Kriegeskorte, N., Kotz, S. A., Gross, J., & Belin, P. (2022). The representational dynamics of perceived voice emotions evolve from categories to dimensions. *Nature Human Behaviour*, *5*.

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*.

Lowe, M. X., Mohsenzadeh, Y., Lahner, B., Charest, I., Oliva, A., & Tengc, S. (2023). Cochlea to categories: The spatiotemporal dynamics of semantic auditory representations. *Cognitive Neuropsychology*, *38*.

Lucia, M. D., Clarke, S., & Murray, M. M. (2010). A temporal hierarchy for conspecific vocalization discrimination in humans. *The Journal of neuroscience*, *30*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2018). Efficient estimation of word representations in vector space. *arXiv*.

Ogg, M., Carlson, T. A., & Slevc, M. R. (2019). The rapid emergence of auditory object representations in cortex reflect central acoustic attributes. *Journal of Cognitive Neuroscience*, *32*.

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*.

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, *114*.

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*.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*.