Temporal dynamics underlying sound discrimination in the human brain

Matthew X. Lowe (mxlowe@mit.edu)
Computer Science and AI Lab (CSAIL), MIT, Cambridge, MA, US

Santani Teng (santani@mit.edu)
Computer Science and AI Lab (CSAIL), MIT, Cambridge, MA, US
Smith-Kettlewell Eye Research Institute, San Francisco, CA, US

Yalda Mohsenzadeh (yalda@mit.edu)
Computer Science and AI Lab (CSAIL), MIT, Cambridge, MA, US

Ian Charest (i.charest@bham.ac.uk)
School of Psychology, University of Birmingham, Birmingham, UK

Dimitrios Pantazis (pantazis@mit.edu)
McGovern Institute for Brain Research, MIT, Cambridge, MA, US

Aude Oliva (oliva@mit.edu)
Computer Science and AI Lab (CSAIL), MIT, Cambridge, MA, US

Abstract

The ability to orient and respond swiftly to our surroundings requires the detection and identification of sounds within moments. Broad descriptors of sounds, such as living and non-living objects, can be discriminated from neural activity starting within the first 100ms of perception, yet when are distinct sound categories (e.g., animals) and individual sounds (e.g., goat) represented and distinguished within the brain across time? To investigate this question, we use magnetoencephalography (MEG) and multivariate analyses of neural activity to examine the time course of audition across individual sounds (e.g., goat) and sound categories (animals, objects, people, spaces). Our results reveal a striking early signal for sound selectivity starting within 80ms after stimulus onset for both individual sounds and sound categories. Sound categories showed a more diffuse generalization across time. Notably, human voices were especially pronounced and distinctive compared with other sound categories. These results illuminate the rapid and parallel emergence of sound identity and category information in the brain, and provide critical evidence that these representations dynamically evolve across time in distinct ways.

Keywords: audition; sound; MEG; multivariate analysis

Introduction

From social cues to environmental threats, the ability to rapidly detect and identify sounds in our environment is critical for efficiently reacting to our surroundings (Teng et al., 2017). Using electroencephalography (EEG), several studies have found differential processing of environmental sound categories for broad category descriptors (living and non-living objects) of sounds beginning just 70ms after stimulus onset (Murray et al., 2006), and distinct processing for human and non-human sounds within 200ms (Capilla et al., 2013; Charest et al., 2009; Delucia et al., 2010). While these investigations have advanced our understanding of the speed of processing in the auditory system, there is still much to be uncovered. For instance, when are distinct sound categories (e.g., animals) and individual sounds (e.g., goat) represented and distinguished within the brain across time? In the current study, we investigate this question using magnetoencephalography (MEG) during an experimental task requiring individuals to listen to sound cues of various stimulus categories (animals, objects, people, spaces). To ensure identification of each sound, participants were asked to form a mental image based on preset descriptors. Multivariate analyses were then used to investigate the time course of audition by tracing the neural signature of auditory processing for individual sounds and sound categories.

Methods and Results

Stimulus Set

Eighty diotic sounds from different semantic sources were selected from a set of 500 natural sounds and divided into four equal categories of twenty sounds each (animals, objects, people, spaces). Sounds were normalized by root mean square values and resampled to 44.1 kHz. Low-level auditory differences across categories were equated by statistically comparing the spectrograms of each sound category with random permutations. Each sound was 500ms in length, including 10ms linear rise and fall times.

Training Procedure

A training procedure was used to familiarize participants with all sounds. Each sound was accompanied with a written description, such as ‘a horse neighing’ (animal), ‘a trumpet playing’ (object), ‘a male shouting angrily’ (people), and ‘howling wind through a city’ (space).

Experimental Paradigm

Eight participants provided informed consent in accordance with guidelines of the MIT Committee on the Use of Humans as Experimental Subjects (COUHES). Participants completed the MEG experimental task in a dimly-lit room and were instructed to keep their eyes closed at all times to prevent artifacts caused by eye movements and blinks. All eighty sounds were randomly presented in each experimental run, and sounds were randomly interleaved with twenty null (no sound) trials and ten oddball (target detection) trials for a total run time of 330 seconds. Each sound trial (including oddball trials) consisted of a 500ms sound followed by 2500ms of silence which preceded the next trial **(Figure 1).** Participants completed 16 experimental runs each. Each run was initiated with a button press, and participants responded to oddball sound detection trials using the same button press.

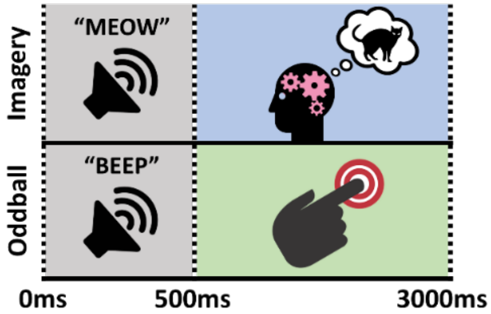

Figure 1. Experimental paradigm. Participants were instructed to form a mental image for each sound presented (above), and press a button when an oddball sound was detected (below).

MEG Acquisition and Preprocessing

Brainstorm software was used to extract trials from 400ms before to 3000ms after target sound onset and preprocess the data. The baseline mean of each sensor was removed and data was smoothed by a low-pass filter of 30Hz.

Statistical Inference

We used nonparametric statistical tests which do not assume any distributions on the data. Our statistical inference on decoding time series were performed by permutation-based cluster-size inference (1000 permutations, 0.05 cluster definition threshold and 0.05 cluster threshold) with null hypothesis of 50% chance level.

MEG Multivariate Pattern Analysis

MEG data was analyzed using multivariate pattern analysis with multivariate noise normalization (Guggenmos et al., 2018). To decode information of the target stimuli, a linear support vector machine (SVM, libsvm implementation) was used as a classifier. In order to reduce computational load, MEG trials of each condition were sub-averaged in groups of 4 with random assignment, resulting in N = 4 pattern vectors per condition. At each time point t of each trial, the MEG data was arranged in a vector of 306 elements. For each pair of sounds and at each time point, the accuracy of SVM classifier was calculated using a leave-one-out procedure. The procedure of sub-averaging and then cross-validation was repeated 100 times. The classifier accuracies were averaged over the repetitions separately for sound pairs. **Figure 2A** shows the decoding time series averaged over pairwise decoding values for all sound categories. Decoding accuracy onsets for both categories and individual sounds occurred at 76ms. **Figure 2B** shows a 80 x 80 representational dissimilarity matrix (RDM) displaying peak MEG decoding at the individual sound level. Pairwise comparisons across categories revealed significant decoding for all categories. Interestingly, higher decoding accuracy was found for human voices compared with other categories.

Visualization with Multidimensional Scaling

To reveal any underlying patterns from MEG decoding matrices, we used multidimensional scaling (MDS) to plot the data into a two-dimensional space such that similar conditions were grouped together and dissimilar conditions far apart. The resulting visualization can be seen in **Figure 2C.**

Temporal Generalization with Multivariate Pattern Analysis

To compare the stability of neural dynamics of sounds, we studied the temporal generalization of their representations by extending the SVM classification procedure (King and Dehaene, 2014). The SVM classifier trained at a given time point t was tested on data at all other time points. The classifier performance in discriminating signals can be

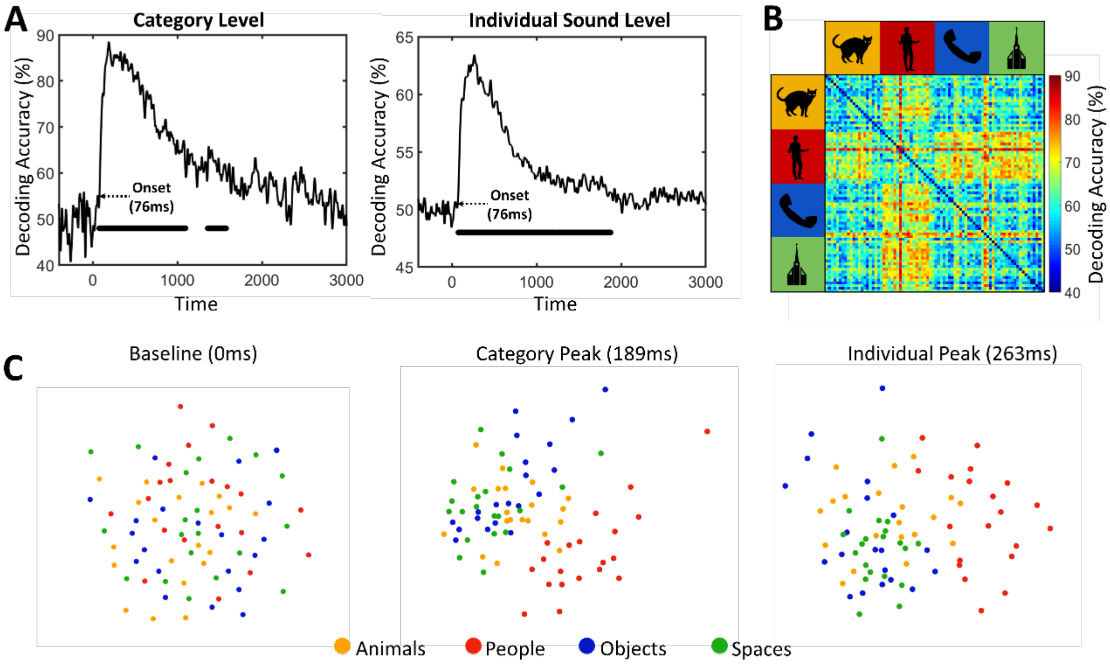

Figure 2. (**A**) Time course of image decoding at the category and individual sound level. Significance statistical tests are with permutations tests using cluster defining thresholds (p < 0.05) shown in black. (**B**) 80 x 80 representational dissimilarity matrix (RDM) displaying peak MEG decoding at the individual sound level (263ms). (**C**) MDS in two-dimensional space at baseline (0ms), peak-latency at the category level (189ms), and peak-latency at the individual sound level (263ms).

generalized to time points with shared representations. This temporal generalization analysis was performed on every pair of images and for each subject. Averaging within sounds and across subjects resulted in a 2D matrix where the x-axis corresponded to training time and y-axis to testing time. The resulting matrices (**Figure 3**), in which each row corresponds to the time (in ms) at which the classifier was trained and each column to the time at which it was tested, reveal a diagonally extended sequence of decoding patterns starting at ~80ms.

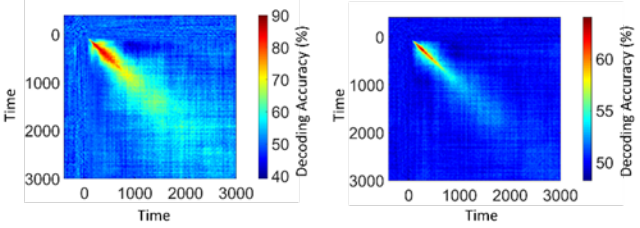

Figure 3. Time course of decoding for sound categories and individual sounds.

The temporal generalization analyses differed between category and individual sound representations. For the individual sound representations, a narrow cluster, closely following the time diagonal, suggests a brief generalization to neighboring time points. At the category level, however, representations showed more diffused significant clusters across time, suggesting that the category representations are maintained consistently over time in the auditory processing stream.

Visualization with Sensorwise Mapping

Results from the temporal generalization analysis suggest a diffuse pattern of decoding across time for category level discriminations of sound. To characterize this distribution within the spatial domain, we conducted a sensorwise analysis of the MEG-response patterns. We generated 102 decoding time courses, one for each sensor triplet location, visualized as sensor maps (**Figure 4).** Statistically significant decoding accuracies were determined via permutation analysis (sign permutation test with 1000 samples, p < 0.05, corrected for FDR across sensor positions at each time point). This analysis revealed primarily bilateral temporal cortical areas contributing to decoding performance during peak-latency decoding (189ms), with decoding extending to bilateral fronto-temporal cortex over time. Sustained distributed activation was found post-stimulus offset (500ms), and bilateral fronto-temporal activity extended past 1000ms post-stimulus onset. These results further confirm a stable cortical representation of sound category over time, and this extended processing should be explored in future investigations.

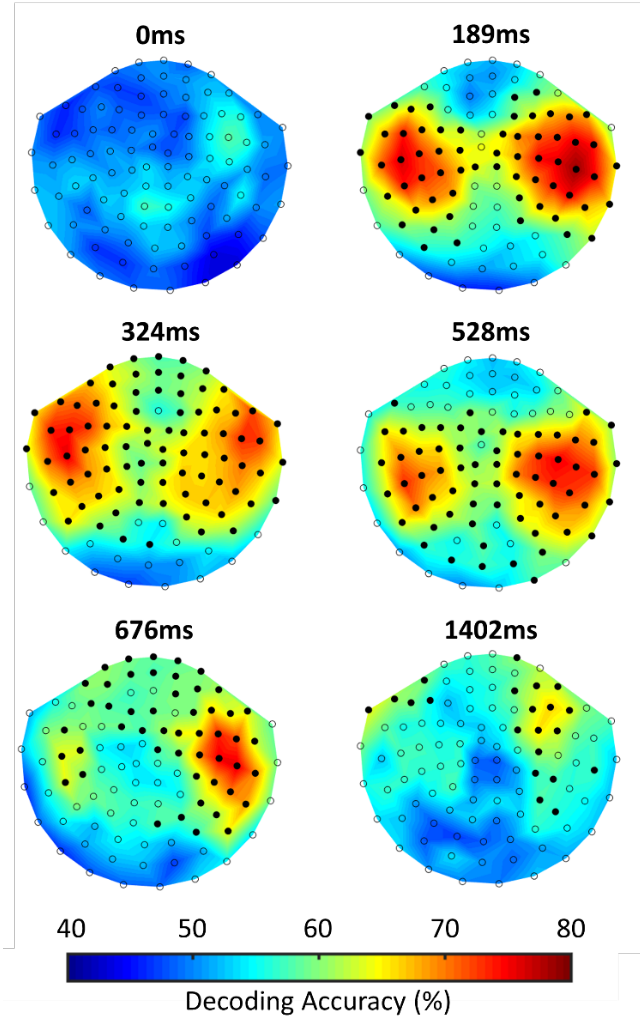

Figure 4. Sensorwise decoding of sound categories for selected time points. Significant decoding is indicated with a black circle over the sensor position ($p < 0.05$; corrected for false discovery rate (FDR) across sensors and time)

Conclusion

Our results provide critical insights into the temporal dynamics underlying auditory processing in the human brain. Notably, we observed a remarkably rapid early signal for both category distinctions and individual sound discriminations starting within the first 80ms of perception, and this signal was sustained even after stimulus offset. Such swift discriminations reflect the ability to recognize and respond to sounds in our environment in mere moments. Sound categories showed a more diffuse and sustained generalization across time, suggesting distinct temporal processing for categories and individual sounds. Representations of human voices were especially distinctive. These results elucidate the differential temporal dynamics for representations of individual sounds and sound categories. A finer understanding of these neural mechanisms provides insights on the cortical representational hierarchy of auditory categorization and thus add temporal computational constraints for modeling human auditory behavior (Aytar et al. 2016; Kell et al. 2018).

Acknowledgements

Vannevar Bush Faculty Fellowship program by the ONR to A.O. (N00014-16-1-3116). Study conducted at the Athinoula A. Martinos Imaging Center, MIBR, MIT.

References

Aytar Y., Vondrick C., Torralba, A.(2016), SoundNet: Learning Sound Representations from Unlabeled Video, NIPS.

Capilla, A., Belin, P., & Gross, J. (2012). The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. *Cerebral Cortex*, *23*(6), 1388-1395.

Charest, I., Perne t, C. R., Rousselet, G. A., Quiñones,I., Latinus, M., Fillion-Bilodeau, S., Chartrand, J.P., & Belin, P. (2009). Electrophysiological evidence for an early processing of human voices. *Bmc Neuroscience*, *10*(1), 127.

De Lucia, M., Clarke, S., & Murray, M. M. (2010). A temporal hierarchy for conspecific vocalization discrimination in humans. *Journal of Neuroscience*, *30*(33), 11210-11221.

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage*, *173*, 434-447.

Kell, A. J., Yamins, D. L., 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*.

King, J. R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203-210.

Murray, M. M., Camen, C., Andino, S. L. G., Bovet, P., & Clarke, S. (2006). Rapid brain discrimination of sounds of objects. *Journal of Neuroscience*, *26*(4), 1293-1302.

Teng, S., Sommer, V. R., Pantazis, D., & Oliva, A. (2017). Hearing Scenes: A Neuromagnetic Signature of Auditory Source and Reverberant Space Separation. *eNeuro*, *4*(1), ENEURO-0007.