Attention During Story Listening Modulates Temporal Receptive Windows Across Human Cortex

Mohammad Shahdloo (shahdloo@ee.bilkent.edu.tr)
Electrical and Electronics Engineering Department,
and National MR Research Center (UMRAM), Bilkent University, Ankara, Turkey

Mert Acar (mert.acar@ug.bilkent.edu.tr)
Electrical and Electronics Engineering Department,
and National MR Research Center (UMRAM), Bilkent University, Ankara, Turkey

Tolga Çukur (cukur@ee.bilkent.edu.tr)
Electrical and Electronics Engineering Department, Neuroscience Program,
and National MR Research Center (UMRAM), Bilkent University, Ankara, Turkey

Abstract

Human brain integrates sensory information across time to represent dynamic complex daily-life environments. Previous studies have shown that higher level perceptual and cognitive cortical areas tend to integrate information over longer time windows, suggesting presence of a hierarchy of temporal receptive windows (TRW) across the brain. Yet, attentional modulations of TRW are unknown. Here, we investigated whether category-based attention modulates TRW within and beyond auditory cortex. Human subjects listened to narrated natural stories and their whole-brain BOLD responses were recorded during three tasks (i.e. passive listening, attention to humans, or attention to places) in separate runs. Contextual representation of the stories derived from an LSTM neural network trained on a language modeling task were used to fit voxelwise encoding models. Contextual information was distorted at multiple time scales to measure TRW during passive listening and during the two attention tasks. Our findings suggest that category-based attention modulates TRW across parietal and frontal cortices. The results also suggest that attention to places extends TRW in parietal cortex.

Keywords: fMRI; natural stories; LSTM; attention; temporal receptive fields.

Introduction

Natural speech is represented at various time scales across the brain, from phonemes to semantically complex sentences (DeWitt & Rauschecker, 2012). Previous studies have introduced the notion of temporal receptive window (TRW) of a cortical circuit as the length of time before a response during which information is integrated and affects that response (Hasson, Yang, Vallines, Heeger, & Rubin, 2008). Moreover, in studies using natural speech stimuli, it has been shown that there exists a hierarchy of increasing TRW from early auditory cortex to higher cognitive areas (Lerner, Honey, Silbert, & Hasson, 2011). Recent reports provided evidence for attentional influences on cortical tuning for low-level features of the auditory stimuli (Mesgarani & Chang, 2012). Yet, it is currently unknown whether attention can modulate TRW.

Here, we investigated this question by studying TRW across the human brain during category-based attention in a natural story listening experiment. Five human subjects listened to over two hours of stories from *The Moth Radio Hour* (Lerner et al., 2011) while performing passive listening or one of the two attention tasks (i.e. attend to "humans", and attend to "places") in different runs. We recorded the whole-brain blood-oxygen-level-dependent (BOLD) responses using functional MRI. We then used rich contextual representations derived from a long short-term memory (LSTM) language model to fit voxelwise encoding models separately for each task and in each individual subject (Jain & Huth, 2018). To estimate TRW, we trained different language models via contextual information that was scrambled at different time scales. Then we fit separate voxelwise encoding models using features obtained from language models trained with different context scrambling levels and assessed prediction performance of models. Finally, we estimated TRW by analyzing the voxelwise prediction performance of the fit models. We compared TRW between the passive listening task and the two attention tasks and assessed attentional sensitivity of TRW across the brain. Furthermore, we compared TRW between the two attention tasks and computed a TRW bias index. Our findings suggest that category-based attention modulates TRW in parietal and frontal cortices. The results also suggest that attention to places extends TRW in parietal cortex. Moreover, TRW in strongly category-selective areas is biased toward the preferred category.

Methods

Experiment Design

The attention experiment was performed in a single sessions consisting of 12 runs. The stimulus consisted of 6 naturally spoken narrative stories from *The Moth Radio Hour* totaling over two hours. A cue word was displayed before each run to indicate the attention task: "humans", or "places". In the attend to humans task, subjects attended to human categories (e.g. woman, man, boy). In the attend to places task, subjects attended to place categories (e.g. building, room, school). The passive listening experiment performed in a single sessions consisting of 10 runs and the stimulus consisted of 10 stories.

MRI Protocols

Data were collected using a 3T Siemens Tim Trio MRI scanner (Siemens Medical Solutions) using a 32-channel receiver coil. Functional data were collected using a T2*-weighted gradient-echo echo-planar-imaging pulse sequence with the following parameters: TR = 2sec, TE = 33msec, water-excitation pulse with flip angle = 70°, voxel size = 2.24mm$\times$2.24mm$\times$4.13mm, field of view = 224mm$\times$224mm, 32 axial slices. To construct cortical surfaces, anatomical data were collected using a three-dimensional T1-weighted magnetization-prepared rapid-acquisition gradient-echo sequence with the following parameters: TR = 2.3 sec, TE = 3.45 msec, flip angle = 10°, voxel size = 1 mm$\times$1 mm$\times$1 mm, field of view = 256 mm$\times$212 mm$\times$256 mm.

Stimulus Embedding

To assess an embedding of the stimulus stories, we used an LSTM model and trained it on a language modeling task (LSTM-LM, (Jain & Huth, 2018)). First, using a large corpus of English text, an embedding space was constructed by computing the co-occurrence statistics between each corpus word and a set of 985 common English words (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016). The corpus consisted of comments scraped from `http://reddit.com`, containing nearly 20M words. Then, the LSTM was trained in the embedding space. The LSTM comprised of 3 layers, each with 985-dimensional hidden states. For each input word, the LSTM-LM used representation of 20 preceding words to output a 985-dimensional representation vector (Jain & Huth, 2018).

Context Scrambling

To assess the scrambled stimulus embedding at level l, we replaced the l^{th} to 20^{th} words in the training samples with random words from the corpus. The LSTM was then trained from scratch using the scrambled context. This procedure led to 20 different stimulus embeddings for $l \in [1;20]$.

Voxelwise Model Fitting and Testing

Voxelwise models were fit using regularized linear regression with an l_2 penalty to avoid overfitting. A nested cross-validation procedure was used to fit model for each voxel. In each of the 20 inner folds, models were fit on the training data for regularization parameters in the range $[2^3;2^{20}]$. Pearsons correlation between actual and predicted responses (prediction score) for the test data were computed. Optimal regularization parameters were then selected to maximize the average prediction score across inner folds. Afterwards, optimized parameters were used to fit models on the union of training and test data in each outer fold. To assess model performance, responses were predicted for the validation data using the fit models. Finally, models and prediction scores for each voxel were averaged across the the 20 outer folds.

Voxelwise Temporal Receptive Windows

In each voxel, we aggregated the prediction scores of the context-scrambled encoding models to form a 20-dimensional prediction profile. To capture the variance in the prediction profiles, we projected the prediction profiles onto the first principal component (PC) of the prediction profiles across all subjects during passive viewing. Only voxels for which the prediction score of the unscrambled model was higher than mean were used to assess PCs. Finally, the 98^{th} percentile of TRWs were adjusted to $[0;1]$.

Sensitivity and bias of TRW

We compared TRW between the passive listening task and the two attention tasks to assess the sensitivity of TRW to category-based attention. For each voxel a sensitivity index was calculated as

$$SI = \frac{1}{2}(|TRW_0 - TRW_H| + |TRW_0 - TRW_P|) \qquad (1)$$

where TRW_0, TRW_H, and TRW_P are the TRW during passive listening, attention to humans, and attention to places. In a voxel with SI of 0 attention does not modulate TRW. A voxel with SI of 1 gets maximally modulated by category-based attention. Finally, we quantified a bias index as

$$BI = TRW_H - TRW_P \qquad (2)$$

Maximized TRW during attention to humans versus attention to places yields positive versus negative BI in the range $[-1;1]$.

Results

We found that TRW increases from early auditory areas toward higher auditory areas and parietal and prefrontal cortices (Fig.1a). The average TRW is $0:47 \pm 0:03$ in early auditory areas (HG, STG, SMG, vPMC, BA44), $0:57 \pm 0:02$ in higher auditory areas (pSTS, BA45), $0:55 \pm 0:02$ in parietal cortex (AG, IPS, SPS, PrC), $0:51 \pm 0:03$ in prefrontal cortex (IFS, MFS, SFG), and $0:50 \pm 0:03$ in category-selective areas in ventral-temporal cortex (FFA, OFA, PPA, RSC). We found that attentional modulation of TRW is relatively low in early auditory areas and category-selective areas (Fig.1b). The average TRW sensitivity index is $0:38 \pm 0:02$ in early auditory areas, $0:41 \pm 0:01$ in higher auditory areas, $0:43 \pm 0:01$ in parietal cortex, $0:41 \pm 0:01$ in prefrontal cortex, and $0:36 \pm 0:05$ in category-selective areas. TRW bias was not significant in early auditory areas in temporal cortex (bootstrap test, $p > 0:05$). However, bias is more prominent in higher cognitive areas in prefrontal cortex (BA45, IFS, MFS, SFG), inferior parietal cortex (IPS and AG), and category-selective areas (Fig.1c). Specifically, the average bias is $-0:04 \pm 0:01$ in inferior parietal cortex, and $-0:02 \pm 0:01$ in prefrontal cortex. The average bias in human-selective areas (FFA and OFA) is $0:03 \pm 0:01$. The average bias in scene-selective areas (PPA and RSC) is $-0:03 \pm 0:02$ (mean±std).

Conclusion

Our results demonstrate that during natural listening, brain optimizes search for categories by integrating information over longer time windows in higher cognitive areas. This finding implies that auditory perception in the real world is facilitated by a mechanism that dynamically modulates temporal receptive windows according to task demand.

Acknowledgments

This work was supported in part by a European Molecular Biology Organization Installation Grant (IG 3028), by a TUBITAK 1001 Grant (118E256), by a BAGEP fellowship, and by a TUBA GEBIP fellowship.

References

DeWitt, I., & Rauschecker, J. P. (2012, February). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(8), E505–E514.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008, March). A Hierarchy of Temporal Receptive Windows in Human Cortex. *The Journal of Neuroscience*, *28*(10), 2539–2550.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016, April). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Jain, S., & Huth, A. G. (2018, January). Incorporating context into language encoding models for fMRI. In *Advances in neural information processing systems* (pp. 6628–6637).

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011, February). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of Neuroscience*, *31*(8), 2906–2915.

Mesgarani, N., & Chang, E. F. (2012, May). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236.