

NATURAL SPEECH REPRESENTATIONS IN THE HUMAN BRAIN DURING A COCKTAIL PARTY

A DISSERTATION SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
NEUROSCIENCE

By
İbrahim Kiremitçi
August 2021

NATURAL SPEECH REPRESENTATIONS IN THE HUMAN
BRAIN DURING A COCKTAIL PARTY

By İbrahim Kiremitçi

August 2021

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Tolga Çukur(Advisor)

Fatoş Tünay Yarman Vural

Hüseyin Boyacı

Hacı Hulusi Kafalgönül

Murat Perit Çakır

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

NATURAL SPEECH REPRESENTATIONS IN THE HUMAN BRAIN DURING A COCKTAIL PARTY

İbrahim Kiremitçi

Ph.D. in NEUROSCIENCE

Advisor: Tolga Çukur

August 2021

Humans are remarkably adept in selectively listening to a desired speaker in a crowded environment, while filtering out non-target speakers in the background. Attention is key to solving this difficult cocktail-party task, yet a detailed characterization of attentional effects on speech representations is lacking. It remains unclear across what levels of speech features and how much attentional modulation occurs in each brain area during the cocktail-party task. Besides, it should be clarified whether unattended speech is represented in cortex during selective listening and if so, at what feature levels its representations are maintained. To address these questions, we recorded whole-brain blood-oxygen-level-dependent (BOLD) responses while subjects either passively listened to single-speaker stories, or selectively attended to a male or a female speaker in temporally-overlaid stories in separate experiments. Spectral, articulatory, and semantic models of the natural stories were constructed to enable comprehensive assessments on the hierarchy of speech features. Intrinsic selectivity profiles were identified via voxelwise models fit to passive listening responses. Attentional modulations were then quantified based on model predictions for attended and unattended stories in the cocktail-party task. We find that acoustic representations are confined to the early auditory cortex whereas linguistic representations are broadly distributed across cortex, that attention causes broad modulations at multiple levels of speech representations (articulatory and semantic) while growing stronger towards later stages of processing, and that unattended speech is represented up to the semantic level in parabelt auditory cortex. These results provide insights on speech perception and attentional mechanisms that underlie the ability to selectively listen to a desired speaker in noisy multi-speaker environments.

Keywords: functional magnetic resonance imaging (fMRI), cocktail-party, dorsal and ventral stream, encoding model, natural speech.

ÖZET

KOKTEYL PARTİSİ SIRASINDA BEYİNDEKİ DOĞAL KONUŞMA TEMSİLLERİ

İbrahim Kiremitçi

NÖROBİLİM, Doktora

Tez Danışmanı: Tolga Çukur

Ağustos 2021

İnsanlar, kalabalık bir ortamda, arka planda hedef olmayan konuşmacıları filtrelerken, istenen bir konuşmacıyı seçici olarak dinleme konusunda oldukça uzmandır. Dikkat, bu zor kokteyl partisi görevini çözmenin anahtarıdır, ancak konuşma temsilleri üzerindeki dikkat etkilerinin ayrıntılı bir karakterizasyonu eksiktir. Kokteyl partisi görevi sırasında her bir beyin bölgesinde hangi seviye konuşma özelliklerinde ve ne kadar dikkat modülasyonunun gerçekleştiği belirsizliğini korumaktadır. Ayrıca, seçici dinleme sırasında dikkat edilmeyen konuşmanın kortekste temsil edilip edilmediği ve temsil ediliyorsa hangi özellik seviyelerinde temsil edildiği netleştirilmelidir. Bu soruları ele almak için, denekler ya pasif olarak tek konuşmacılı hikayeleri dinlerken ya da üst üste bindirilmiş hikayelerdeki erkek ya da kadın konuşmacıdan birine seçici olarak dikkat kesilmişken, tüm beyin kan-oksijen-seviyesi-bağımlı (BOLD) yanıtları ayrı deneylerde kaydedildi. Doğal hikayelerin spektral, artikulator ve semantik modelleri, konuşma özelliklerinin hiyerarşisi hakkında kapsamlı değerlendirmeler yapabilmek için oluşturuldu. İçsel seçicilik profilleri, pasif dinleme yanıtlarına fit edilmiş voksel modeller aracılığıyla belirlendi. Daha sonra, kokteyl partisi görevindeki dikkat edilen ve edilmeyen öyküler için model tahminlerine dayalı olarak dikkat modülasyonları hesaplandı. Akustik temsillerin erken işitsel korteksle sınırlı olduğunu, oysa linguistik temsillerin korteks boyunca geniş bir şekilde dağıldığını; dikkatin konuşma hiyerarşisinin sonraki aşamalarına doğru giderek güçlenen bir şekilde, konuşma temsillerinin birden çok seviyesinde (artikulator ve semantik) geniş modülasyonlara neden olduğunu, ve dikkat edilmeyen konuşmanın parabelt işitsel kortekste semantik seviyeye kadar temsil edildiğini bulduk. Bu sonuçlar, konuşma algısını ve gürültülü, çok konuşmacılı ortamlarda istenen bir konuşmacıyı seçici olarak dinleme yeteneğinin altında yatan dikkat mekanizmalarını anlamamıza büyük katkı sağlamaktadır.

Anahtar sözcükler: fonksiyonel manyetik rezonans görüntüleme (fMRI), kokteyl

partisi, dorsal ve ventral yolak, kodlama modeli, doğal konuşma.



Acknowledgement

While writing this last and most satisfying part of my thesis, I feel very grateful to be able to complete this difficult and transforming journey. It was invaluable to take any kind of support from people who care and love you throughout this journey, and I feel very appreciated for their loving kindness.

First of all, I want to express my gratitude to my supervisor Tolga Çukur for his patience, support and efforts, that enabled me to complete this journey. Writing a science article was the most difficult but also the most instructive experience in my life, and that was thanks to his natural talent in teaching this invaluable skill. Growing as an independent researcher requires a sensitive balance between autonomy and taking guidance, and he was superior at providing both at the right time and place.

I would like to thank my TİK members Fatoş Yarman Vural and Hüseyin Boyacı for their support and flexibility during TİK meetings. It was hard to pursue a PhD while working in the industry, and their kindness made everything easier for me. I thank my jury member Hulusi Kafalıgönül for his friendly and encouraging conversations in UMRAM. I also would like to express my sincere gratitude to my jury member Murat Perit Çakır to accept being involved in important milestones of this journey.

Next, I would like to thank ICON Lab members, with special thanks to Ümit, Salman, Emin, Mohammad and Özgür. Having very helpful, collaborative, approachable and friendly lab mates made everything easier and cheerful. I also would like to express my appreciation to all UMRAM family for creating such an inspiring working environment.

My dear friends Burcu Yargıçoğlu and Emine Çimen are the ones who planted the seeds in my mind to start this journey, and they always made me feel their strong wish and support not to give up. My dear friend Selami Çiftçi has always been very inspiring with his very positive attitude in challenging times. Murat Yaman has been a great companion making my Bilkent days more fun. I would like to thank them all for being my friends and for touching my life with their distinct and special colors.

I feel very lucky to have great friends whose love, companionship, support, and

kindness have always been my shelter throughout this journey. I want to express my sincerest gratitude to my beloved friends Özge, Pınar, Serdar, Göksel, Nihan, Kemal, Ertuğrul, Gamze, Burcu, Ceren, and Melike. Each of them has a great contribution to this journey with their special colors; and I cannot imagine my life without any of them.

Finally, I would like to express my deepest gratitude to my family for their unconditional love, unshakable belief, and infinite support. Their love and care have always been my pole star when I feel lost, and their belief has always been my strength when I feel weak. I feel very grateful to be born in such a great family full of compassion and love. I dedicate this thesis to my family.

Contents

1	Introduction	1
1.1	Publications	4
1.2	Outline	5
2	Background	6
2.1	Magnetic Resonance Imaging (MRI)	6
2.2	functional Magnetic Resonance Imaging (fMRI)	9
2.3	Naturalistic Stimuli and Encoding Models	10
2.4	A Brief History of the Cocktail-party Problem	13
3	Attentional Modulation of Hierarchical Speech Representations in a Multi-Talker Environment	16
3.1	Introduction	16
3.2	Materials and Methods	20
3.2.1	Participants	20
3.2.2	Stimuli	20
3.2.3	Experimental procedures	21
3.2.4	MRI data collection and preprocessing	23
3.2.5	Visualization on cortical flatmaps	24
3.2.6	ROI definitions and abbreviations	24
3.2.7	Model construction	25
3.2.8	Analyses	30
3.2.9	Statistical tests	36
3.3	Results	37
3.3.1	Intrinsic cortical representations of natural speech	37

3.3.2	Attentional modulation of multi-level speech representations	48
3.3.3	Cortical representation of unattended speech	57
3.4	Discussion	60
3.4.1	Speech representations	60
3.4.2	Attentional modulations	61
3.4.3	Representation of the unattended speech	64
4	Conclusion	65
4.1	Future Directions	66
A	Supplementary Materials	92
A.1	Supplementary Tables	93
A.2	Supplementary Figures	97

List of Figures

- 2.1 **Encoding models.** Brain can be considered as a non-linear system that maps the stimuli to the neural activity. Encoding models put this non-linear system into a framework with three abstract spaces: input, feature, and activity spaces. Each stimulus sample lays as a point in the input space. Each point in the input space is first nonlinearly mapped to a point in the feature space, which corresponds to a latent representation of the stimulus. Then, a linear mapping associates each point in the feature space to the certain voxel activities, which correspond to the points in the activity space. Figure reprinted from Neuroimage, 56(2), T. Naselaris, K.N. Kay, S. Nishimoto, J.L. Gallant, “Encoding and decoding in fMRI”, Page 11, Copyright (2011), with permission from Elsevier. 12
- 2.2 **Cognitive Theories for Selective Attention.** Three stood-out theories for explaining the role of selective attention during cocktail-party: early-selection filter theory, attenuation theory, and late-selection theory. 14

3.1 **Experimental design.** **a.** *Passive-listening experiment.* 10 stories from Moth-Radio-Hour were used to compile a single-speaker stimulus set. Subjects were instructed to listen to the stimulus vigilantly without any explicit task in the passive-listening experiment. **b.** *Cocktail-party experiment.* A pair of stories told by individuals of different genders were selected from the single-speaker stimulus set and overlaid temporally to generate a two-speaker stimulus set. Subjects were instructed to attend either to the male or female speaker in the cocktail-party experiment. The same two-speaker story was presented twice in separate runs while the target speaker was varied. Attention condition was fixed within runs and it alternated across runs. 20

3.2 **Multi-level speech features.** Three distinct feature spaces were constructed to represent natural speech at multiple levels: spectral, articulatory, and semantic spaces. Speech waveforms were projected separately on these spaces to form stimulus matrices. The spectral feature matrix captured the cochleogram features of the stimulus in 93 channels having center frequencies between 115 and 9920 Hz. The articulatory feature matrix captured the mapping of each phoneme in the stimulus to 22 binary articulation features. The semantic feature matrix captured the statistical co-occurrences of each word in the stimulus with 985 common words in English. Each feature matrix was Lanczos-filtered at a cut-off frequency of 0.25 Hz and downsampled to 0.5 Hz to match the sampling rate of fMRI. Natural speech might contain intrinsic stimulus correlations among spectral, articulatory, and semantic features. To prevent potential biases due to stimulus correlations, we decorrelated the three feature matrices examined here via Gram-Schmidt orthogonalization (see Methods). The decorrelated feature matrices were used for modeling BOLD responses. 26

3.3 Caption next page 31

3.3 **Modeling procedures.** **a.** *Voxelwise modeling.* Voxelwise models were fit in individual subjects using passive-listening data. To account for hemodynamic response, a linearized four-tap finite impulse response (FIR) filter spanning delayed effects at 2-8 sec was used. Models were fit via L2-regularized linear regression. BOLD responses were predicted based on fit voxelwise models on held-out passive-listening data. Prediction scores were taken as the Pearson’s correlation between predicted and measured BOLD responses. For a given subject, speech-selective voxels were taken as the union of voxels significantly predicted by spectral, articulatory, or semantic models ($q(FDR) < 10^{-5}, t - test$). **b.** *Assessment of attentional modulation.* Passive-listening models for single voxels were tested on cocktail-party data to quantify attentional modulations in selectivity. In a given run, one of the speakers in a two-speaker story was attended while the other speaker was ignored. Separate response predictions were obtained using the isolated story stimuli for the attended speaker and for the unattended speaker. Since a voxel can represent information from both attended and unattended stimuli, a linear combination of these predicted responses was considered with varying combination weights (wc in $[0\ 1]$). BOLD responses were predicted based on each combination weight separately. Three separate prediction scores were calculated based on only the attended stimulus ($wc = 1$), based on only the unattended stimulus ($wc = 0$), and based on the optimal combination of the two stimuli. A model-specific attention index, (AI_m) was then computed as the ratio of the difference in prediction scores for attended versus unattended stories to the prediction score for their optimal combination (see Methods). 32

3.4 Caption next page 38

3.4 **Prediction scores of voxelwise speech models.** **a.** *Spectral model.* Prediction scores for voxels significantly predicted by the spectral model are plotted on the flattened cortical surface of subject S4. Medial and lateral views of the inflated hemispheres are also shown above and below the flatmaps, respectively. Colors indicate the value of the noise-corrected prediction scores (see legend). White lines encircle ROIs that are found based on an automatic atlas-based cortical parcellation. Labels of some relevant ROIs are shown (see Methods for ROI abbreviations). In this subject, significantly predicted voxels by the spectral model lie mainly in the dorsal aspect of the superior temporal cortex. **b.** *Articulatory model.* Significantly predicted voxels by the articulatory model are located mainly in the dorsal and lateral aspects of the superior temporal cortex extending to STS, and in inferior frontal regions. **c.** *Semantic model.* Significantly predicted voxels by the semantic model are broadly distributed across cortex covering most of the temporal, prefrontal and parietal cortical regions except the early auditory cortex. 39

3.5 Caption next page 40

- 3.5 Selectivity for multi-level speech features. a. *Model-specific selectivity indices.*** Single-voxel prediction scores on passive-listening data were used to quantify the selectivity of each ROI to underlying model features. Model-specific prediction scores were averaged across speech-selective voxels within each ROI and normalized such that the cumulative score from all models was 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory, and semantic models (mean $\pm$ sem across subjects). Significant indices are marked with * ($p < 0.05$; see Supp. Figs. A.1-5 for selectivity indices of individual subjects). ROIs in perisylvian cortex are displayed (see Figure 3.6 for non-perisylvian ROIs; see Methods for ROI abbreviations). ROIs in LH and RH are shown in the top and bottom panels, respectively. POP_R and PreG_R that did not have consistent speech selectivity in individual subjects were excluded (see Methods). **b. *Intrinsic selectivity profiles.*** Selectivity profiles of cortical ROIs averaged across subjects are shown on the cortical flatmap of a representative subject (S4). Significant articulatory, semantic, and spectral selectivity indices of each ROI are projected to the red, green, and blue channels of the RGB colormap (see Methods). This analysis only included ROIs with consistent selectivity for speech features in each individual subject. Medial and lateral views of the inflated hemispheres are also shown. A progression from low-intermediate to high-level speech representations are apparent across bilateral temporal cortex in the superior-inferior direction; consistently in all subjects (see Supp. Fig. A.6 for selectivity profiles of individual subjects). Meanwhile, semantic selectivity is dominant in many higher-order regions within the parietal and frontal cortices (bilateral AG, IPS, SPS, PrC, PCC, POS, PTR, IFS, SFS, SFG, MFG and left POP) ($p < 0.05$; see Supp. Figs. A.1-5). These results support the view that speech representations are hierarchically organized across cortex with partial overlap between spectral, articulatory and semantic representations in early to intermediate stages of auditory processing..

3.6 Model-specific selectivity indices for non-perisylvian ROIs.

Single-voxel prediction scores on passive-listening data were used to quantify the degree of selectivity of each ROI to the underlying model features under passive listening. Prediction scores for a given model were averaged across speech-selective voxels within each ROI, and then normalized such that the cumulative prediction score from all models summed to 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory and semantic models (mean $\pm$ sem across subjects). Significant indices are marked with *, colored according to the model ($p < 0.05$, bootstrap test; see Supp. Figs. A.1-5 for selectivity indices of individual subjects). Only ROIs in non-perisylvian cortex are displayed. **a.** ROIs in LH. **b.** ROIs in RH. mOTS_R that did not have consistent speech selectivity in individual subjects was excluded (see Methods). These results show that selectivity is primarily semantic in most of the non-perisylvian ROIs. 42

3.7 **Representational complexity.** To characterize the complexity of speech representations, a complexity index, (*CI*), was taken as the relative tuning of an ROI for low- versus high-level speech features. *CI* is in the range of [0, 1], where higher values indicate stronger tuning for semantic features and lower values indicate stronger tuning for spectral features. Bar plots indicate *CI* (mean $\pm$ sem across subjects). Indices that are consistently significant in each individual subject are marked with * ($p < 10^{-4}$, bootstrap test).

a. Perisylvian ROIs. ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. Early auditory regions such as HG/HS have *CI* close to 0, whereas higher-order regions like PTR have *CI* tending to 1. These results indicate that more complex features are represented in higher-order regions compared to the earlier ones.

b. Non-perisylvian ROIs. ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. Higher-order regions in non-perisylvian cortex mostly have similar representational complexity that is close to 1. 43

- 3.8 **Cortical hierarchy of representational complexity.** **a.** *Complexity index and auditory pathways.* To visualize how representational complexity distributes across cortex, of cortical ROIs averaged across subjects are displayed on the flatmap of a representative subject (S4; see legend). To illustrate the gradients in across the hierarchy of speech processing, dorsal and ventral pathways are shown with blue and green lines, respectively: left dorsal-1 (LD-1), left dorsal-2 (LD-2) and right dorsal (RD), left ventral-1 (LV-1), left ventral-2 (LV-2), right ventral-1 (RV-1) and right ventral-2 (RV-2). Squares mark regions where pathways begin; arrows mark regions where pathways end; and circles mark relay regions in between. **b.** *Complexity hierarchies.* Bar plots display (mean $\pm$ sem across subjects) along LD-1, LD-2, RD, LV-1, LV-2, RV-1 and RV-2, shown in separate panels. Significant differences in CI between consecutive ROIs are marked with brackets ($p < 0.05$, bootstrap test). Significant gradients in CI are: $CI_{HG/HS} < CI_{PT} < CI_{POP}$ in LD-1, $CI_{HG/HS} < CI_{PT} < CI_{PreG}$ in LD-2, $CI_{HG/HS} < CI_{mSTG} < CI_{mSTS}$ in LV-1 and RV-1, and $CI_{HG/HS} < CI_{mSTG} < CI_{aSTG}$ in LV-2 and RV-2. These results suggest that hierarchical speech representations are systematically organized in multiple gradients across cortex. 45

3.9 **Predicting cocktail-party responses.** Passive-listening models were tested during the cocktail-party task by predicting BOLD responses in the cocktail-party data. Since a voxel might represent information from both attended and unattended stimuli, response predictions were expressed as a convex combination of individual predictions for the attended and unattended story within each two-speaker story. Prediction scores were computed as the combination weights (w_c) were varied in $[0\ 1]$ (see Methods). Prediction scores for a given model were averaged across speech-selective voxels within each ROI (r_c). The normalized scores of spectral, articulatory, and semantic models are displayed in several representative ROIs (HG/HS, HG/HS, and PT). Solid and dashed lines indicate mean and %95 confidence intervals across subjects. Scores based on only the attended story (r_a), based on only the unattended story (r_u), and based on the optimal combination of the two (r_{max}) are marked with circles. For the *spectral model* in left HG/HS, r_{max} is larger than r_a ($p < 10^{-4}$); and the optimal combination equally weighs attended and unattended stories. For the *articulatory model* in left HG/HS, r_a is larger than r_u ($p < 10^{-4}$), while r_{max} is greater than r_a ($p < 10^{-2}$). Besides, the optimal combination puts slightly higher weight to attended story than unattended story. For the *semantic model* in left PT, r_a is much higher than r_u ($p < 10^{-4}$), and the optimal combination puts much greater weight to attended story than unattended one. These representative results imply that attention may have divergent effects at various levels of speech representations across cortex. 49

3.10 Caption next page 51

3.10 Attentional modulation of multi-level speech representations.

a. Model-specific attention indices. A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI in perisylvian cortex, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across subjects), and their sum yields the overall modulation (see Figure 3.11 for non-perisylvian ROIs). Significantly positive indices are marked with * ($p < 0.05$, bootstrap test; see Supp. Figs. A.7-11 for attention indices of individual subjects). ROIs in the LH and RH are shown in top and bottom panels, respectively. These results show that selectivity modulations distribute broadly across cortex at the linguistic level (articulatory and semantic).

b. Attentional modulation profiles. Modulation profiles averaged across subjects are displayed on the flattened cortical surface of a representative subject (S4). Significantly positive articulatory, semantic, and spectral attention indices are projected onto the red, green and blue channels of the colormap (see Methods). A progression in the level of speech representations dominantly modulated is apparent from HG/HS to MTG across bilateral temporal cortex (see Supp. Fig. A.12 for modulation profiles of individual subjects). Articulatory modulation is dominant in one end of the dorsal stream (left PreG), whereas semantic modulation becomes dominant in both ends of the ventral stream (bilateral PTR and MTG) ($p < 0.05$; see Supp. Figs. A.7-11 and A.12). On the other hand, semantic modulation is dominant in most of the higher-order regions in the parietal and frontal cortices consistently in all subjects (bilateral AG, SPS, PrC, PCC, POS, SFG, SFS, and PTR; left MFG; and right IPS) ($p < 0.05$; see Supp. Figs. A.7-11).

3.11 **Model-specific attention indices for non-perisylvian ROIs.**

A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI in non-perisylvian cortex, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across subjects), and their sum yields the overall modulation. Significant indices are marked with * ($p < 0.05$, bootstrap test; see Supp. Figs. A.7-11 for attention indices of individual subjects). **a. ROIs in LH.** Modulations in left IPS and mOTS are not consistently significant in all subjects ($p > 0.05$). **b. ROIs in RH.** mOTS_R that did not have consistent speech selectivity in individual subjects was excluded (see Methods). These results indicate that selectivity modulations manifest primarily at the semantic level in most of the non-perisylvian ROIs. 53

3.12 Caption next page 55

3.12 **Global attentional modulation.** **a.** *Global attention index.* To quantify overall modulatory effects on selectivity across all examined feature levels, global attentional modulation (*gAI*) was computed by summing spectral, articulatory, and semantic attention indices (see Methods). *gAI* is in the range of [-1,1] and a value of zero indicates no modulation. Colors indicate significantly positive *gAI* averaged across subjects (see legend; see Supp. Fig. A.13 for bar plots of *gAI* across cortex). Dorsal and ventral pathways are shown with blue and green lines, respectively: left dorsal-1 (LD-1), left dorsal-2 (LD-2) and right dorsal (RD), left ventral-1 (LV-1), left ventral-2 (LV-2), right ventral-1 (RV-1) and right ventral-2 (RV-2). Squares mark regions where pathways begin; arrows mark regions where pathways end; and circles mark relay regions in *gAI* between. **b.** *Modulation hierarchies.* Bar plots display *gAI* (mean $\pm$ sem across subjects) along LD-1, LD-2, RD, LV-1, LV-2, RV-1 and RV-2, shown in separate panels. Significant differences in between consecutive ROIs are marked with brackets ($p < 0.05$, bootstrap test; see Supp. Fig. A.14 for single-subject results). Significant gradients in *gAI* are: $gAI_{PT} < gAI_{SMG} < gAI_{POP}$ in LD-1, $gAI_{PT} < gAI_{SMG} < gAI_{PreG}$ in LD-2, $gAI_{HG/HS} < gAI_{mSTG} < gAI_{mSTS}$ in LV-1, $gAI_{HG/HS} < gAI_{mSTG} < gAI_{aSTG}$ in LV-2, and $gAI_{mSTG} < gAI_{aSTG}$ in RV-2. In the left hemisphere, *gAI* gradually increases from early auditory regions to higher-order regions across the dorsal and ventral pathways. Similar patterns are also observed in the right hemisphere, although the gradients in *gAI* are less consistent across subjects. 56

3.13 Caption next page 58

3.13 Representation of unattended speech. Passive-listening models were tested on cocktail-party data to assess representation of unattended speech during the cocktail-party task. Prediction scores were calculated separately for a combination model comprising features of both attended and unattended stories (r_{max} : optimal convex combination) and an individual model only comprising features of the attended story (r_a). Significant difference in prediction between the two models is an indication that BOLD responses carry significant information on unattended speech. Bar plots display normalized prediction scores (mean $\pm$ sem across subjects; combination model in light gray and individual model in gray). Significant scores are marked with * ($p < 10^{-4}$, bootstrap test; see Supp. Figs. A.15-19 for single-subject results), and significant differences are marked with brackets ($p < 0.05$). Prediction scores are displayed for ROIs in the dorsal and ventral streams, with significant selectivity for given model features. **a. Left hemisphere.** Spectral representations of unattended speech extend up to PT across the dorsal stream ($HG/HS \rightarrow PT \rightarrow SMG \rightarrow (POP, PreG)$) and are constrained to HG/HS across the ventral stream ($HG/HS \rightarrow mSTG \rightarrow mSTS \rightarrow MTG$) and ($HG/HS \rightarrow mSTG \rightarrow aSTG \rightarrow PTR$). Articulatory representations of unattended speech extend up to PT across the dorsal stream and are constrained to HG/HS across the ventral stream. No significant semantic representation is apparent. **b. Right hemisphere.** Spectral representations of unattended speech extend up to SMG across the dorsal stream and are constrained to HG/HS across the ventral stream. Articulatory representations of unattended speech extend up to PT across the dorsal stream, and up to mSTG across the ventral stream. Semantic representations are found only in mSTS. These results suggest that processing of unattended speech is not constrained at spectral level but extends to articulatory and semantic level.

A.1 Model-specific selectivity indices for subject S1. Single-voxel prediction scores on passive-listening data were used to quantify the degree of selectivity of each ROI to the underlying model features under passive listening. Prediction scores for a given model were averaged across speech-selective voxels within each ROI, and then normalized such that the cumulative prediction score from all models summed to 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory and semantic models (mean $\pm$ sem across ROI's speech-selective voxels). Significant indices are marked with *, colored according to the model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a. Perisylvian ROIs.** ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. . . 98

A.2 Model-specific selectivity indices for subject S2. Single-voxel prediction scores on passive-listening data were used to quantify the degree of selectivity of each ROI to the underlying model features under passive listening. Prediction scores for a given model were averaged across speech-selective voxels within each ROI, and then normalized such that the cumulative prediction score from all models summed to 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory and semantic models (mean $\pm$ sem across ROI's speech-selective voxels). Significant indices are marked with *, colored according to the model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a.** *Perisylvian ROIs.* ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b.** *Non-perisylvian ROIs.* ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. . . 99

A.3 Model-specific selectivity indices for subject S3. Single-voxel prediction scores on passive-listening data were used to quantify the degree of selectivity of each ROI to the underlying model features under passive listening. Prediction scores for a given model were averaged across speech-selective voxels within each ROI, and then normalized such that the cumulative prediction score from all models summed to 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory and semantic models (mean $\pm$ sem across ROI's speech-selective voxels). Significant indices are marked with *, colored according to the model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a.** *Perisylvian ROIs.* ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b.** *Non-perisylvian ROIs.* ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. . . 100

A.4 Model-specific selectivity indices for subject S4. Single-voxel prediction scores on passive-listening data were used to quantify the degree of selectivity of each ROI to the underlying model features under passive listening. Prediction scores for a given model were averaged across speech-selective voxels within each ROI, and then normalized such that the cumulative prediction score from all models summed to 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory and semantic models (mean $\pm$ sem across ROI's speech-selective voxels). Significant indices are marked with *, colored according to the model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a. Perisylvian ROIs.** ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. . . 101

A.5 **Model-specific selectivity indices for subject S5.** Single-voxel prediction scores on passive-listening data were used to quantify the degree of selectivity of each ROI to the underlying model features under passive listening. Prediction scores for a given model were averaged across speech-selective voxels within each ROI, and then normalized such that the cumulative prediction score from all models summed to 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory and semantic models (mean $\pm$ sem across ROI's speech-selective voxels). Significant indices are marked with *, colored according to the model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a.** *Perisylvian ROIs.* ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b.** *Non-perisylvian ROIs.* ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. . . . 102

A.6 **Intrinsic selectivity profiles.** Selectivity profiles of perisylvian ROIs in subjects S1-S5 are shown on the cortical flatmaps. Significant articulatory, semantic, and spectral selectivity indices are projected onto the red, green, and blue channels of the RGB colormap (see Methods). **a.** *Left perisylvian ROIs.* **b.** *Right perisylvian ROIs.* Consistently across subjects, a progression from low- and intermediate-level to high-level speech representations are apparent across bilateral temporal cortex in superior-inferior direction ($HG/HS \rightarrow mSTG \rightarrow mSTS \rightarrow MTG$). These results support the view that speech representations are hierarchically organized across processing hierarchy of speech with partial overlap between spectral, articulatory and semantic representations in early to intermediate stages of auditory processing. 103

A.7 **Model-specific attention indices for subject S1.** A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across speech-selective voxels in each ROI), and their sum yields the overall modulation. Significantly positive modulations are marked with *, colored according to the corresponding model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a.** *Perisylvian ROIs.* ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b.** *Non-perisylvian ROIs.* ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. 104

A.8 **Model-specific attention indices for subject S2.** A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across speech-selective voxels in each ROI), and their sum yields the overall modulation. Significantly positive modulations are marked with *, colored according to the corresponding model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a.** *Perisylvian ROIs.* ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b.** *Non-perisylvian ROIs.* ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. 105

A.9 **Model-specific attention indices for subject S3.** A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across speech-selective voxels in each ROI), and their sum yields the overall modulation. Significantly positive modulations are marked with *, colored according to the corresponding model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a.** *Perisylvian ROIs.* ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b.** *Non-perisylvian ROIs.* ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. 106

A.10 **Model-specific attention indices for subject S4.** A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across speech-selective voxels in each ROI), and their sum yields the overall modulation. Significantly positive modulations are marked with *, colored according to the corresponding model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a.** *Perisylvian ROIs.* ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b.** *Non-perisylvian ROIs.* ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. 107

A.11 **Model-specific attention indices for subject S5.** A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across speech-selective voxels in each ROI), and their sum yields the overall modulation. Significantly positive modulations are marked with *, colored according to the corresponding model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a.** *Perisylvian ROIs.* ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b.** *Non-perisylvian ROIs.* ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. 108

A.12 **Attentional modulation profiles for individual subjects.** Modulation profiles of perisylvian ROIs in subjects S1-S5 are shown on the cortical flatmaps. Significantly positive articulatory, semantic, and spectral attention indices were projected to the red, green, and blue channels of the RGB colormap (see Methods). **a.** *Left perisylvian ROIs.* **b.** *Right perisylvian ROIs.* In all subjects, a progression in the level of speech representations dominantly modulated is apparent across bilateral temporal cortex in the superior-inferior direction ($HG/HS \rightarrow mSTG \rightarrow mSTS \rightarrow MTG$). Semantic modulation is dominant at both ends (MTG and PTR) of bilateral ventral stream, while articulatory modulation is dominant at one end (PreG) of left dorsal stream ($p < 0.05$, bootstrap test; see also Supp. Figs. A.7-11). There is no consistent tendency towards dominant articulatory or semantic modulation at the other end of the left dorsal stream (POP). 109

- A.13 Global attention indices.** To quantify overall modulatory effects on selectivity across all examined feature levels, global attentional modulation (gAI) was computed by summing spectral, articulatory, and semantic attention indices (see Methods). gAI is in the range of $[-1,1]$, where a positive index indicates attentional modulation in favor of the attended stimuli and a negative index indicates modulation in favor of the unattended stimuli. A value of zero indicates no modulation. Bar plots indicate gAI (mean $\pm$ sem across subjects). Significant indices are marked with * ($p < 0.05$, bootstrap test; see the bar plots in Supp. Figs. A.7-11 for gAI in individual subjects). **a. Perisylvian ROIs.** ROIs within perisylvian cortex in LH and RH are displayed in top and bottom panels, respectively. gAI is significant in all perisylvian ROIs consistently in all subjects except in right PT and SMG ($p > 0.05$). **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. gAI is significant in all non-perisylvian ROIs consistently in all subjects except in left IPS and mOTS ($p > 0.05$). 110
- A.14 Modulation hierarchies for individual subjects.** **a. Modulation hierarchies for subject S1.** Gradients in gAI across LD-1, LD-2, RD, LV-1, LV-2, RV-1 and RV-2 are shown in subfigures labeled accordingly. Bar plots display gAI in each ROI (mean $\pm$ sem across speech-selective voxels in each ROI). Only ROIs within a given trajectory are included in the corresponding subfigure. Significant differences in gAI between consecutive ROIs along the trajectory are marked with brackets ($p < 0.05$, bootstrap test). **b. Modulation hierarchies for subject S2. c. Modulation hierarchies for subject S3. d. Modulation hierarchies for subject S4. e. Modulation hierarchies for subject S5.** Gradients consistently obtained in each individual subject are ($p < 0.05$): $gAI_{PT} < gAI_{SMG} < gAI_{POP}$ in LD-1, $gAI_{PT} < gAI_{SMG} < gAI_{PreG}$ in LD-2, $gAI_{HG/HS} < gAI_{mSTG} < gAI_{mSTS}$ in LV-1, $gAI_{HG/HS} < gAI_{mSTG} < gAI_{aSTG}$ in LV-2, and $gAI_{mSTG} < gAI_{aSTG}$ in RV-2. 111

A.15 Representation of unattended speech in subject S1.

Passive-listening models were tested on cocktail-party data to assess representation of unattended speech during the cocktail-party task. Prediction scores were calculated separately for a combination model comprising features of both attended and unattended stories (r_{max} : optimal convex combination) and an individual model only comprising features of the attended story (r_a). Significant difference in prediction between the two models is an indication that BOLD responses carry significant information on unattended speech. Bar plots display normalized prediction scores (mean $\pm$ sem across speech-selective voxels in each ROI; combination model in light gray and individual model in gray). Significant scores are marked with * ($p < 10^{-4}$, bootstrap test), and significant differences are marked with brackets ($p < 0.05$). Prediction scores are only displayed for ROIs in the dorsal and ventral streams, with significant selectivity for given model features. **a.** *Left hemisphere.* Spectral, articulatory and semantic representations are labeled in subplots accordingly. **b.** *Right hemisphere.* Spectral, articulatory and semantic representations are labeled in subplots accordingly. . 112

A.16 Representation of unattended speech in subject S2.

Passive-listening models were tested on cocktail-party data to assess representation of unattended speech during the cocktail-party task. Prediction scores were calculated separately for a combination model comprising features of both attended and unattended stories (r_{max} : optimal convex combination) and an individual model only comprising features of the attended story (r_a). Significant difference in prediction between the two models is an indication that BOLD responses carry significant information on unattended speech. Bar plots display normalized prediction scores (mean $\pm$ sem across speech-selective voxels in each ROI; combination model in light gray and individual model in gray). Significant scores are marked with * ($p < 10^{-4}$, bootstrap test), and significant differences are marked with brackets ($p < 0.05$). Prediction scores are only displayed for ROIs in the dorsal and ventral streams, with significant selectivity for given model features. **a.** *Left hemisphere.* Spectral, articulatory and semantic representations are labeled in subplots accordingly. **b.** *Right hemisphere.* Spectral, articulatory and semantic representations are labeled in subplots accordingly. . 113

A.17 Representation of unattended speech in subject S3.

Passive-listening models were tested on cocktail-party data to assess representation of unattended speech during the cocktail-party task. Prediction scores were calculated separately for a combination model comprising features of both attended and unattended stories (r_{max} : optimal convex combination) and an individual model only comprising features of the attended story (r_a). Significant difference in prediction between the two models is an indication that BOLD responses carry significant information on unattended speech. Bar plots display normalized prediction scores (mean $\pm$ sem across speech-selective voxels in each ROI; combination model in light gray and individual model in gray). Significant scores are marked with * ($p < 10^{-4}$, bootstrap test), and significant differences are marked with brackets ($p < 0.05$). Prediction scores are only displayed for ROIs in the dorsal and ventral streams, with significant selectivity for given model features. **a.** *Left hemisphere.* Spectral, articulatory and semantic representations are labeled in subplots accordingly. **b.** *Right hemisphere.* Spectral, articulatory and semantic representations are labeled in subplots accordingly. . 114

A.18 Representation of unattended speech in subject S4.

Passive-listening models were tested on cocktail-party data to assess representation of unattended speech during the cocktail-party task. Prediction scores were calculated separately for a combination model comprising features of both attended and unattended stories (r_{max} : optimal convex combination) and an individual model only comprising features of the attended story (r_a). Significant difference in prediction between the two models is an indication that BOLD responses carry significant information on unattended speech. Bar plots display normalized prediction scores (mean $\pm$ sem across speech-selective voxels in each ROI; combination model in light gray and individual model in gray). Significant scores are marked with * ($p < 10^{-4}$, bootstrap test), and significant differences are marked with brackets ($p < 0.05$). Prediction scores are only displayed for ROIs in the dorsal and ventral streams, with significant selectivity for given model features. **a.** *Left hemisphere.* Spectral, articulatory and semantic representations are labeled in subplots accordingly. **b.** *Right hemisphere.* Spectral, articulatory and semantic representations are labeled in subplots accordingly. . 115

A.19 Representation of unattended speech in subject S5.

Passive-listening models were tested on cocktail-party data to assess representation of unattended speech during the cocktail-party task. Prediction scores were calculated separately for a combination model comprising features of both attended and unattended stories (r_{max} : optimal convex combination) and an individual model only comprising features of the attended story (r_a). Significant difference in prediction between the two models is an indication that BOLD responses carry significant information on unattended speech. Bar plots display normalized prediction scores (mean $\pm$ sem across speech-selective voxels in each ROI; combination model in light gray and individual model in gray). Significant scores are marked with * ($p < 10^{-4}$, bootstrap test), and significant differences are marked with brackets ($p < 0.05$). Prediction scores are only displayed for ROIs in the dorsal and ventral streams, with significant selectivity for given model features. **a.** *Left hemisphere.* Spectral, articulatory and semantic representations are labeled in subplots accordingly. **b.** *Right hemisphere.* Spectral, articulatory and semantic representations are labeled in subplots accordingly. . 116

List of Tables

3.1	Phoneme-articulatory features map	28
3.2	Hemispheric asymmetries in CI and gAI.	47
A.1	Inter-run head motion in the cocktail-party experiment	94
A.2	Number of significantly predicted voxels within ROIs in the left hemisphere	95
A.3	Number of significantly predicted voxels within ROIs in the right hemisphere	96

Chapter 1

Introduction

The brain is the most complex and mysterious structure in the universe with about 86-billion neurons and tremendous number of connections between neurons [1]. Although it comprises only 2% of the body mass, it consumes a significant proportion of the total body energy (about 20%) to control a broad range of sensory, motor and cognitive functions from the most primitive like smelling to most complex like problem solving and planning [2]. Among all these functions, language is one of the most unique abilities of the human brain [3]. Thus, how the brain manifests such a complex cognitive function has been an intriguing question in the field of neuroscience for a long time. The relationship between brain damage and certain speech disorders dates back to the writings of Hippocrates (400 BC); however, one of the most important scientific contributions to the neural basis of language was in the 19th century [4]. Through postmortem examination of a patient with intact comprehension but impaired articulation, Broca was the first who proposed that posterior part of the inferior frontal gyrus is the locus in charge of articulation [5]. This region was later called Broca's Area. About a decade later, Wernicke proposed posterior part of the superior temporal gyrus as the region related to comprehension through post-mortem examination of two patients with intact articulation but impaired comprehension [5]. This area was later called Wernicke's Area. He further proposed a language model which suggested that Broca's area and Wernicke's area are the regions where motor and

sensory aspects of the language emerge, respectively, and that these two regions are interconnected because sensory representations of speech are also necessary for articulation [6]. This simple model became the foundation of the classic model of language neurobiology and was the prevailing paradigm for a long time. With the development of new technology and methods to study the brain, the growth in the field of language neurobiology accelerated. Especially, with the advent of magnetic resonance imaging (MRI), the studies in that field exploded. Through structural MRI, the relationship between various language deficits and lesions in different brain regions could be examined in vivo non-invasively with great spatial resolutions [7, 8]. Through functional MRI (fMRI), different aspects of language from acoustic to semantic began to be studied using very smart and controlled experimental paradigms and stimuli [9, 10, 11]. Through the diffusion tensor imaging capability of MRI, white matter connections between different language-related areas started to be explored in detail to understand the information flow during language perception [12, 13, 14, 15, 16]. Based on the studies conducted with these new emerging neuroimaging capabilities, the classic model has been refined and improved; and even new models have been proposed [17, 18, 19, 20]. All these developments have put MRI in the spotlight as a very useful and versatile tool to study different aspects of language perception in the brain.

The only thing more fascinating than this unique ability of the human brain is the capability to use this skill effortlessly even in very distracting and noisy environments. Imagine that you are at a cocktail party surrounded by people in small groups who are chatting about totally different topics, gesturing, laughing and making a lot of noise. Foods and drinks are being served. Glasses are clinking. In the background, a band or speaker is playing music. In all this cacophony of sound, you are talking with someone about the latest rumors in the office. Suddenly, you hear your name being spoken in the neighboring group and with an irresistible curiosity, you start eavesdropping on them while continuing your conversation. This routine scenario for the human brain becomes a very complex problem when looking from an engineering perspective. First, all the auditory signals from different sources will add with each other and arrive at your ear as a single auditory signal. Then, the target speech signal should be segregated

from this mixture by your early auditory system. Phonemes/syllables, words, and sentences will need to be formed hierarchically from this acoustic signal for its content to be comprehended. While this target speech keeps being processed, the rest of the mixture should continue to be processed as well to capture any salient or important information in the background. There is still no technology which comes close to the excellent auditory performance of the human brain even today. On the other hand, it is quite crucial to understand how the brain accomplishes this to develop technologies for aiding people with hearing impairments [21, 22, 23, 24]. This topic has been drawing the attention of many researchers for a long time; and Cherry was the first who described this phenomenon as the cocktail-party problem [25]. In a series of experiments using a dichotic listening paradigm, he asked the subjects to repeat aloud the passage that was spoken into one ear while ignoring the other. By manipulating the stimuli spoken into the ignored ear at acoustic or linguistic level, he examined the extent of the level that the passage in the ignored ear was processed through post-experiment questionnaire. In that way, Cherry set an experimental framework to study cocktail-party problem behaviorally. With the increasing popularity of fMRI for studying different aspects of language perception, Cherry’s framework was borrowed by many neuroscientists to understand how brain resolves the cocktail-party problem [26, 27]. Most of these studies used controlled stimuli such as simple tones, isolated syllables or words. However, a more natural setting is required to gain a more thorough understanding of the cocktail-party phenomenon.

Fortunately, in the last decade, with improvements in data acquisition and computational resources, there has been a growing interest in using computational frameworks that leverage brain data collected during hour-long exposure to naturalistic stimuli [28, 29]. One standout framework is encoding modeling. Within this framework, a linear mapping between the features of the stimuli and evoked cortical responses is estimated [30]. In this way, cortical responses for a novel stimulus can be predicted using the estimated models. Significant predictions in a cortical area imply the representations of the stimuli features in this area. An important advantage of this framework is its flexibility to test many hypotheses related to the brain function being studied with less experimentation

than conventional methods [30]. In case of language perception, conventional methods need to conduct a series of experiments using controlled stimuli ranging from tones or rotated speech to words to explore different levels of language processing from acoustic to semantic. A major drawback of this approach is that it is ineffective and expensive as more experimentation is required. Another one is that the collected data is strongly dependent on the hypothesis being tested; therefore, it does not allow the reuse of the same data to test new hypotheses in the future. Lastly and more importantly, such designs assume that the low-level cognitive processes involved in naturalistic language perception are the same as those involved in low-level auditory stimulus perception. However, this assumption ignores the fact that human brain is more experienced in processing naturalistic speech than any other auditory stimuli [31]. These compelling reasons motivate using encoding models and naturalistic stimuli to research the neural basis of language perception.

1.1 Publications

The content of this thesis has been partially presented and published in the following conferences and journals:

- I. Kiremitci, O. Yilmaz, A.G. Huth, U. Keles, T. Cukur. Attentional modulation of articulatory and semantic representations during multi-speaker natural speech perception. 138.15. 2019 Neuroscience Meeting Planner. Chicago: Society for Neuroscience, 2019. Online.
- I. Kiremitçi, Ö. Yilmaz, E. Çelik, M. Shahdloo, A.G. Huth, T. Çukur. Attentional Modulation of hierarchical speech representations in a multi-talker environment. bioRxiv, 2020.
- I. Kiremitçi, Ö. Yilmaz, E. Çelik, M. Shahdloo, A.G. Huth, T. Çukur. Attentional Modulation of hierarchical speech representations in a multitalker environment. Cerebral Cortex, 2021. Online.

1.2 Outline

The rest of this thesis is organized as follows. In chapter 2, we give some background information about MRI, fMRI, encoding models, and cocktail-party problem. In chapter 3, we use encoding modeling framework to investigate multi-level speech representations during a passive-listening task and attentional modulations of these representations during a cocktail-party task. Finally, in chapter 4 we provide concluding remarks and directions for future work.

Chapter 2

Background

In the last a few decades, there has been a tremendous growth in the field of neuroscience, mainly because of the advances in the neuroimaging modalities including electroencephalography (EEG), magnetoencephalography (MEG), and fMRI. Their safe and non-invasive nature made them quite popular to study a wide range of sensory, motor and cognitive functions of the human brain. The greater temporal resolution of EEG/MEG modalities make them suitable to study temporal aspects of many brain functions [32]. On the other hand, superior spatial resolution of fMRI has made it quite attractive to be used for anatomical mapping of brain functions. With the cooperation of the naturalistic stimuli and powerful computational methods like encoding models, fMRI has been a preferred modality to decipher the codes of language cognition in many recent studies. This chapter will introduce the reader some background information for the main principles of fMRI and encoding models.

2.1 Magnetic Resonance Imaging (MRI)

MRI is an imaging modality that is used to acquire 3-dimensional images of the biological tissues non-invasively. An MRI machine consists of three main

components: a superconducting electromagnet for creating a static strong magnetic field; gradient coils for creating spatial variation in the magnetic field; and radio-frequency (RF) coils for transmitting and receiving signals at the resonance frequency of the volume being imaged [33]. The main principle of the measured MRI signal is based on the spinning property of the protons in Hydrogen atoms. Protons can be thought as tiny spinning magnets [34]. These spins are randomly oriented in the absence of any external magnetic field. When a static magnetic field is applied, they all align with the direction of the external field. This alignment creates a net magnetization. Therefore, when a live is put into an MRI machine, the strong static magnetic field creates a net longitudinal magnetization in the direction of the external field due to the aligned Hydrogen protons of that live [33]. This magnetization is dependent on the strength of the static field. When an RF pulse is transmitted through RF coils to a particular slice of the volume being imaged – resonant with the spinning frequency of the protons in that slice –, the net magnetization of that slice is transferred to the transverse plane (transverse magnetization; [34]). This transverse magnetization is measured through the RF receiver coil. However, the measured signal is the net magnetization of the whole measured slice. To be able to estimate the spatial contributions to the measured signal, gradient coils are used. These coils create spatial gradients in the magnetic field applied to the slice to enable spatial encoding of the measured signal [28]. By controlling these gradients in a parametric way, multiple measurements are taken from the same slice to form raw MRI data, namely the k-space data. Finally, an inverse Fourier Transformation is performed on the k-space data to reconstruct image reflecting signal intensity for each volume element, namely voxels.

MRI modality is highly versatile to emphasize different tissue characteristics and create contrasts by adjusting scanning parameters like how often we transmit RF pulse sequence (TR) and how soon after transmission we start to collect data (TE) [33]. When RF pulse is applied and removed, the spins come back to the equilibrium which transfers the net magnetization back to the longitudinal plane. Therefore, the longitudinal magnetization increases (longitudinal relaxation) while transverse magnetization decays (transverse relaxation) in time.

However, the duration for the longitudinal magnetization reaching to its equilibrium (T1), and transverse magnetization decaying to zero (T2) vary from tissue to tissue. The differences in longitudinal and transversal relaxation times between tissues are two important sources of the contrasts. Therefore, controlling TR and TE values could determine whether T1 or T2 dominate the contrasts in the images (T1-weighted or T2-weighted contrasts). High-resolution structural image of the brain is mainly obtained through these contrasts. Besides, better contrast can be obtained through dynamic adjustments of these parameters with more optimal pulse sequences [35, 36, 37, 38, 39, 40]

An important concern during MRI data acquisition is the trade-off between scan times and the spatial resolution of the reconstructed images. Higher spatial resolution basically requires more measurements in k-space. However, that causes longer scan times. On the other hand, it is crucial to accelerate the image acquisition without sacrificing the spatial resolution to exploit the full potential of MRI modality in clinical uses. Hardware related improvements like using multiple RF coils to transmit and receive signal from multiple slices of a volume in parallel accelerated image acquisition dramatically [41, 42, 33]. Innovative image reconstruction algorithms accompanied to these multi-coil and multi-acquisition solutions [43, 44, 45, 46, 47]. On the other hand, taking less k-space measurements and reconstructing image from this undersampled data provided an effective and cheaper alternative for accelerating acquisition [48, 49, 50, 51, 52]. In the last decade, many novel methods have been proposed for successful image reconstruction from undersampled data including compress sensing [53, 54, 55, 56, 57, 58, 59, 60, 61] and deep neural networks [62, 63, 64, 65, 66, 67, 68, 69]

2.2 functional Magnetic Resonance Imaging (fMRI)

Functional MRI adds time as a fourth dimension to the 3-dimensional MR images, thus providing information about the changes in voxel intensities over time. fMRI is mainly based on local inhomogeneities in the magnetic field. Intrinsic defects in the magnet itself can cause such inhomogeneities, resulting in shorter transverse relaxation times ($T2^*$) [70]. MRI acquisition sequence can also be adjusted to be sensitive to dynamics of local cerebral blood flow, which is another source of local inhomogeneities in the field [28]. That is mainly because of the different magnetic properties of oxygenated and deoxygenated hemoglobin in the blood [71]. Deoxygenated hemoglobin becomes paramagnetic. This creates more distortion in the magnetic field, resulting in more local inhomogeneity [33]. Consequently, transverse magnetization decays even faster, resulting in a lower signal. On the other hand, oxygenated hemoglobin decreases the inhomogeneities in the field, resulting in a higher signal. The dynamic changes in the ratio of oxygenated to deoxygenated hemoglobin in the blood is one of the sources of contrast in the measured signal. fMRI mainly uses this blood-oxygen-level-dependent (BOLD) contrast [34].

Brain hemodynamics are driven by the neural activity. Regional cerebral blood flow and glucose utilization increase with neural activation; and this increase is greater than oxygen consumption rate [34, 72, 73]. The net effect of neural excitation is therefore an increase in oxyhemoglobin concentration and a consequent stronger MRI signal [34]. Thus, measured BOLD signal reflects changes in the neural activity. The alteration in BOLD signal induced by neural activity is called hemodynamic response function (HRF) [33]. A canonical form of HRF is proposed with certain characteristics: an initial dip in BOLD signal within the first 1-2 seconds after the onset of neural activity; a following increase in signal beginning within 1-2 seconds and peaking within 5-8 seconds; a following decrease in signal below a baseline level, sustaining about 10 seconds (post-stimulus undershoot); and a return to the baseline within 15-20 seconds after the onset of

neural activity [74, 75, 33]. However, the timing and shape of the HRF can differ across the brain, within an individual and across individuals [76, 33]. To overcome this handicap, a finite impulse response model has been used for a more flexible estimation of HRF in recent studies [77, 78, 79].

The collected fMRI data does not contain only BOLD signal but also several noise components caused by hardware limitations and the subjects [33]. It is crucial to preprocess data to emphasize only BOLD signal, before relating it to the underlying neural activity. One source of noise is thermal motion of electrons within the subject and the scanner; this thermal noise is random and can be removed through spatial smoothing of fMRI signal across a few voxels [32]. Another source of noise is the scanner drift. This drift causes a low-frequency trend in fMRI signal; and it can be removed through a temporal filter [33]. There are also physiological noises caused by head motion, respiration and heartbeat of the subjects [32, 80]. To assure the measured signal to come from the same brain region, motion is corrected through realignment of the fMRI images using rigid-body transformation [33]. However, motion-correction does not completely remove its effects on fMRI signal. To regress out its contribution from the signal, estimated motion parameters can be used as nuisance covariates during analysis. To remove noise caused by heartbeat and respiration, these physiological signals can be first measured during scanning. Like motion parameters, these measurements can then be used as nuisance covariates during analysis [33].

2.3 Naturalistic Stimuli and Encoding Models

To leverage full potential of fMRI as a modality to study the brain, it is very crucial to combine optimal experimental design, stimuli and analysis method. In the early fMRI studies, the main approach was to use a block design which separates experimental conditions into blocks [33]. The signal measured at a specific condition is contrasted against a control condition to find related activity. To maximize the contrasts, the same stimulus is given repeatedly in blocks lasting at least a few tens of seconds. However, such block designs do not leverage the

temporal resolution of the fMRI which is on the scale of a few seconds [75]. Event-related design emerged as an alternative to the block design. In that design, the subjects are exposed to short discrete events, with randomized timings [33]. In that way, the timing of the activations could be better estimated; and the psychological effects including boredom, habituation or fatigue could be minimized [32]. However, in both conventional paradigms, the main tendency is to use simpler tasks with unnatural stimuli to explore brain functions. This approach deduces complex high-level processes into simpler ones with a few controllable dimensions. Therefore, many important aspects and their interactions remain ignored or undetected.

There has been an increasing trend to employ naturalistic stimuli and tasks in recent fMRI studies [81, 82, 29, 83, 84, 28, 85]. Sophisticated statistical methods with advanced computational power enable investigating the relationship between rich information embedded in the naturalistic stimuli and the BOLD responses [28]. One standout framework is encoding modeling. In this framework, the input stimulus is associated with a set of voxel activities through a latent feature representation of the stimulus (Figure 2.1; [30]). For this, each point in the input space is first mapped -possibly nonlinearly- to a point in a feature space [30, 28]. Then, each point in the feature space is mapped linearly to a point in the activity space, where BOLD responses lay [30, 28]. To give an example, imagine that you have a speech signal. For extracting spectrogram features from this signal, you estimate the power density of that signal at specific frequency bands. Associating the input signal to amplitude values at certain frequency bands is a non-linear mapping. This non-linear mapping can be based on cochlear operations to make a more biologically plausible mapping from sound to spectrogram features. Then, to find most relevant regions whose activities are directly driven by these spectrogram features, a linear relation is assumed between the spectrogram features and the brain activity [30]. A linear encoding model is fit to do this mapping.

A point in the feature space does not always correspond to the same point in the activity space. Instead, many repetitions of the same stimulus sample create a cloud of points in the activity space [30]. That is mainly because of the noise in the fMRI signal, that is not relevant to the task or stimuli. The size

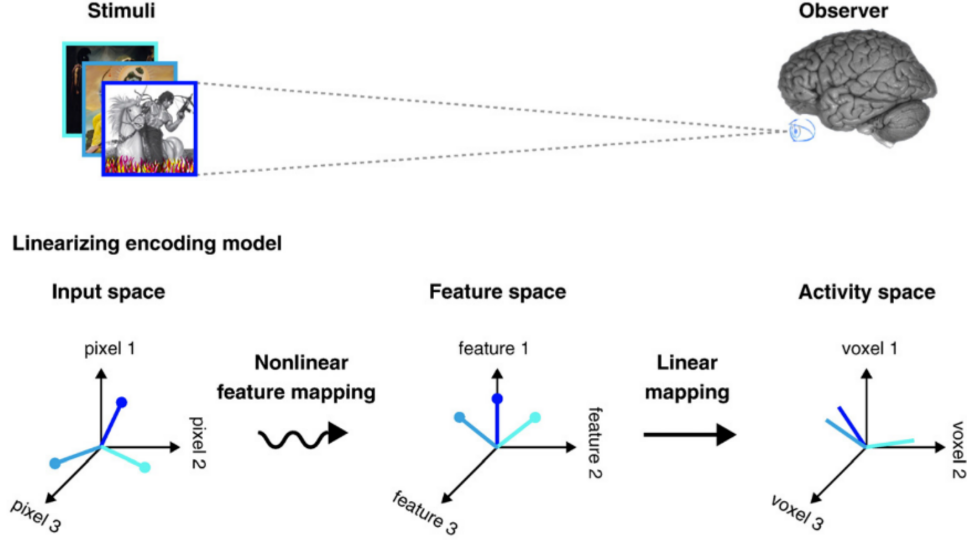


Figure 2.1: **Encoding models.** Brain can be considered as a non-linear system that maps the stimuli to the neural activity. Encoding models put this non-linear system into a framework with three abstract spaces: input, feature, and activity spaces. Each stimulus sample lays as a point in the input space. Each point in the input space is first nonlinearly mapped to a point in the feature space, which corresponds to a latent representation of the stimulus. Then, a linear mapping associates each point in the feature space to the certain voxel activities, which correspond to the points in the activity space. Figure reprinted from Neuroimage, 56(2), T. Naselaris, K.N. Kay, S. Nishimoto, J.L. Gallant, “Encoding and decoding in fMRI”, Page 11, Copyright (2011), with permission from Elsevier.

and shape of the cloud can be determined by an encoding distribution, $p(r|f(s))$, where s is a stimulus sample, $f(s)$ is the corresponding feature point, and r is the measured activity across a set of voxels [30]. This encoding distribution gives the probability of an activity pattern r , given the features $f(s)$; and the best the encoding model can do is to identify the most likely point in the cloud [30]. Therefore, the ideal encoding model can be expressed as:

$$W^T f(s) = \operatorname{argmax}_r p(r|f(s)), \quad (2.1)$$

where W is the model weights that linearly map points in the feature space to the activity space [30]. For fMRI data, the encoding distribution can be assumed as a multivariate Gaussian distribution [30]:

$$p(r|f(s)) \sim \exp(-0.5(r - W^T f(s))^T \Sigma^{-1}(r - W^T f(s))) \quad (2.2)$$

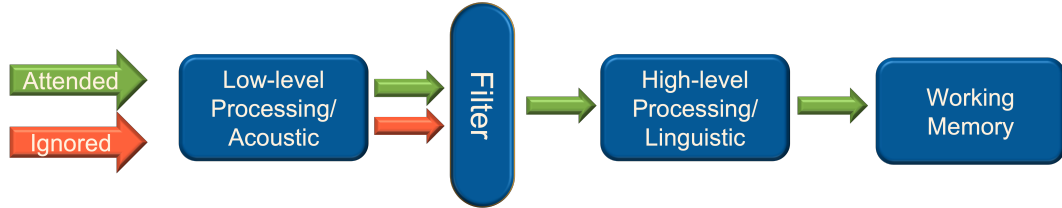
In this case, the model weights can be obtained by solving a regularized regression problem [28]. We leverage this linear encoding modeling framework in Chapter 3 to model BOLD responses evoked by natural stories.

2.4 A Brief History of the Cocktail-party Problem

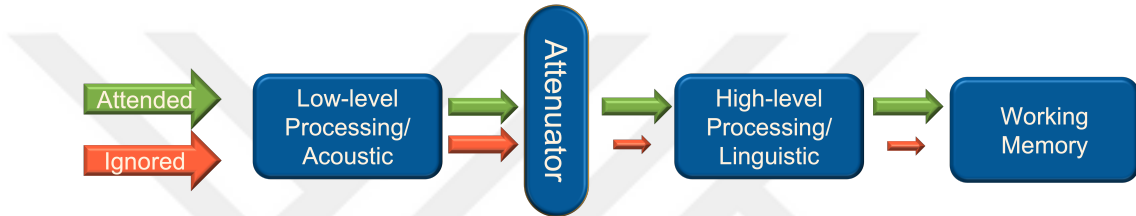
How brain can recognize a specific speech among many has intrigued many researchers for a long time. Cherry was the first who defined this question as “cocktail-party problem” [25]. To investigate this unique ability of the brain, he conducted several experiments. In these experiments, he employed a dichotic listening task where subjects were given different auditory streams to different ears and asked to attend one of the streams [25]. To ensure the subjects’ attention, he also asked them to repeat what they hear in the attended ear (shadowing). The subjects could successfully repeat what they heard in the attended ear but not in the unattended ear. To investigate the extent of the processing of unattended speech, Cherry also changed some acoustic and linguistic attributes of the unattended stream and asked subjects whether they recognized these changes. While subjects noticed changes in acoustic attributes like transition from male to female speaker or from regular speech to simple tones, they failed to detect changes in linguistic attributes like transition from English to German, or from regular to reversed speech. Based on his findings, Cherry suggested that only low-level features of the unattended speech are processed [25].

In the next decades, many researchers designed smart experiments based on Cherry’s experimental paradigm to investigate cocktail-party problem in detail [86, 87, 88, 89, 90]. In this early era of the cocktail-party problem, three cognitive theories stood out to explain findings of the behavioral studies as shown in Figure

1. Early-Selection Filter Theory (Broadbent 1958)



2. Attenuation Theory (Treisman 1960)



3. Late-Selection Theory (Deutsch and Deutsch 1963)

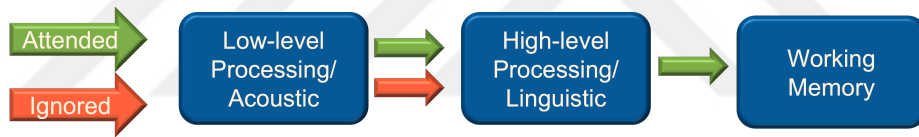


Figure 2.2: **Cognitive Theories for Selective Attention.** Three stood-out theories for explaining the role of selective attention during cocktail-party: early-selection filter theory, attenuation theory, and late-selection theory.

2.2. Early-selection filter theory [91] states that both attended and unattended streams are processed at acoustic level; however, there is a filter at an early stage of processing, which passes only the attended stream for further processing. This theory was very consistent with Cherry's findings and explained well how only the changes in low-level attributes of the unattended stream were recognized by the subjects in his experiments. However, it fell short in explaining how subjects could hear some salient words like their own names in the unattended stream [86] or how some words in the unattended stream could be spoken when they were contextually relevant to the attended stream [87]. Therefore, Treisman proposed the attenuation theory [87], which suggested an attenuator in the early stage of processing instead of a filter. That attenuator passes both the attended and unattended streams for further processing; however, unattended stream is weakened compared to the attended one. When the words are very salient or when they are primed by the context of the attended stream, these weaker signals

can be sufficient to be recognized. Finally, in the late selection theory [88], it was proposed that both attended and unattended stream are further processed at high-level; however, only attended stream enters working memory. Some studies using indirect measures of the unattended stream such as increase in galvanic skin response for the conditioning words in the unattended stream also supported this theory [89].

With the advances in non-invasive technologies, neuroscience community also started to investigate how human brain resolves this cocktail-party problem. Hillyard et al. [92] conducted the first electrophysiological study (EEG) with human, employing a dichotic listening task with simple tones. They found enhanced event-related potential (ERP) at around 100 ms for the tones in the attended ear. Many studies with controlled stimuli repeatedly found enhanced cortical responses to the attended stream in the following years [93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 26]. These studies consistently suggest an attentional gain mechanism in favor of the attended stream. However, it is only recently possible to investigate neural aspects of this problem in detail by simulating more realistic cocktail-party environments thanks to the dramatic increase in computational power and the consequent usage of more advanced analysis methods including encoding models.

Chapter 3

Attentional Modulation of Hierarchical Speech Representations in a Multi-Talker Environment

3.1 Introduction

Humans are highly adept at perceiving a target speaker in crowded multi-speaker environments [21, 106, 107]. Auditory attention is key to behavioral performance in this difficult “cocktail-party problem” [25, 108, 23, 109, 22]. Literature consistently reports that attention selectively enhances cortical responses to the target stream in auditory cortex and beyond, while filtering out non-target background streams [110, 111, 93, 112, 26, 94, 95, 96, 101, 102, 113, 114]. However, the precise link between the response modulations and underlying speech representations is less clear. Speech representations are hierarchically organized across multiple stages of processing in cortex, with each stage selective for diverse information ranging from low-level acoustic to high-level semantic features [115, 116, 17, 18, 19, 117, 20, 118, 119, 120]. Thus, a principal question is to

what extent attention modulates these multi-level speech representations in the human brain during a cocktail-party task [27, 121].

Recent electrophysiology studies on the cocktail-party problem have investigated attentional response modulations for natural speech stimuli [122, 123, 124, 125, 126, 127, 128, 129, 130, 131]. Ding and Simon [123, 124] fit spectrotemporal encoding models to predict cortical responses from the speech spectrogram. Attentional modulation in the peak amplitude of spectrotemporal response functions was reported in planum temporale in favor of the attended speech. Mesgarani and Chang [125] built decoding models to estimate the speech spectrogram from responses measured during passive listening and examined the similarity of the decoded spectrogram during a cocktail-party task to the isolated spectrograms of attended versus unattended speech. They found higher similarity to attended speech in non-primary auditory cortex. Zion Golumbic et al. [127] reported amplitude modulations in speech-envelope response functions towards attended speech across auditory, inferior temporal, frontal and parietal cortices. Other studies using decoding models have similarly reported higher decoding performance for the speech envelope of the attended stream in auditory, prefrontal, motor and somatosensory cortices [128, 131, 129] further identified peak amplitude response modulations for sub-lexical features including word onset and cohort entropy in temporal cortex. Note that because these electrophysiology studies fit models for acoustic or sub-lexical features, the reported attentional modulations primarily comprised relatively low-level speech representations.

Several neuroimaging studies have also examined whole-brain cortical responses to natural speech in a cocktail-party setting [132, 133, 134, 135, 136, 137, 138]. In the study of Hill and Miller [135], subjects were given an attention cue (attend to pitch, attend to location or rest) and later exposed to multiple speech stimuli where they performed the cued task. Partly overlapping frontal and parietal activations were reported, during both the cue and the stimulus exposure periods, as an effect of attention to pitch or location in contrast to rest. Furthermore, pitch-based attention was found to elicit higher responses in bilateral posterior and right middle superior temporal sulcus, whereas location-based attention elicited higher responses in left intraparietal sulcus. In alignment

with electrophysiology studies, these results suggest that attention modulates relatively low-level speech representations comprising paralinguistic features. In a more recent study, Regev et al. [137] measured responses under two distinct conditions: while subjects were presented bimodal speech-text stories and asked to attend to either the auditory or visual stimulus, and while subjects were presented unimodal speech or text stories. Inter-subject response correlations were measured between unimodal and bimodal conditions. Broad attentional modulations in response correlation were reported from primary auditory cortex to temporal, parietal and frontal regions in favor of the attended modality. While this finding raises the possibility that attention might also affect representations in higher-order regions, a systematic characterization of individual speech features that drive attentional modulations across cortex is lacking.

An equally important question regarding the cocktail-party problem is whether unattended speech streams are represented in cortex despite the reported modulations in favor of the target stream [109, 27]. Electrophysiology studies on this issue identified representations of low-level spectrogram and speech envelope features of unattended speech in early auditory areas [123, 124, 125, 127, 128, 129, 131], but no representations of linguistic features [129]. Meanwhile, a group of neuroimaging studies found broader cortical responses to unattended speech in superior temporal cortex [139, 140, 136, 141, 142, 137]. Specifically, Wild et al. [136] and Evans et al. [142] reported enhanced activity associated with the intelligibility of unattended stream in parts of superior temporal cortex extending to superior temporal sulcus. Although this implies that responses in relatively higher auditory areas carry some information regarding unattended speech stimuli, the specific features of unattended speech that are represented across the cortical hierarchy of speech is lacking.

Here we investigated whether and how attention affects representations of attended and unattended natural speech across cortex. To address these questions, we systematically examined multi-level speech representations during a diotic cocktail-party task using naturalistic stimuli. Whole-brain BOLD responses were recorded in two separate experiments (Figure 3.1) while subjects were presented engaging spoken narratives from The Moth Radio Hour. In the passive-listening

experiment, subjects listened to single-speaker stories for over two hours. Separate voxelwise models were fit that measured selectivity for spectral, articulatory, and semantic features of natural speech during passive listening [119]. In the cocktail-party experiment, subjects listened to temporally-overlaid speech streams from two speakers while attending to a target category (male or female speaker). To assess attentional modulation in functional selectivity, voxelwise models fit during passive listening were used to predict responses for the cocktail-party experiment. Model performances were calculated separately for attended and unattended stories. Attentional modulation was taken as the difference between these two performance measurements. Comprehensive analyses were conducted to examine the intrinsic complexity and attentional modulation of multi-level speech representations and to investigate up to what level of speech features unattended speech is represented across cortex.

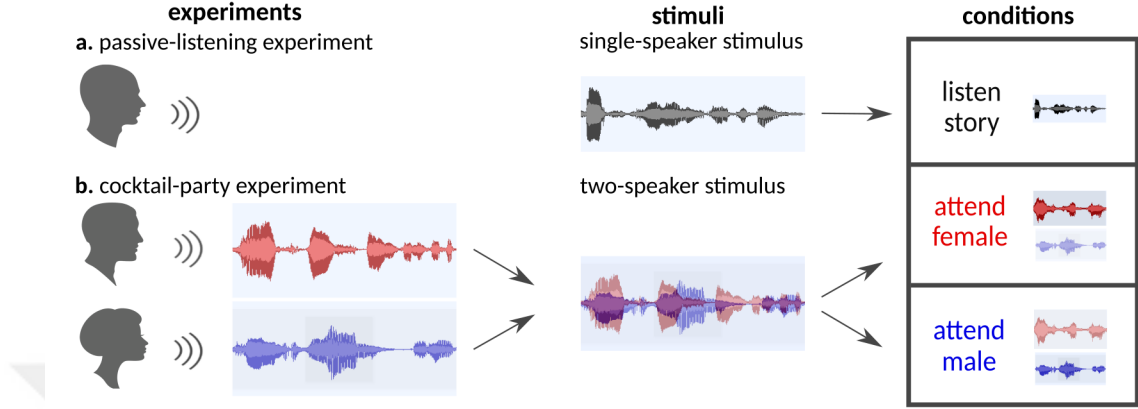


Figure 3.1: **Experimental design.** **a.** *Passive-listening experiment.* 10 stories from Moth-Radio-Hour were used to compile a single-speaker stimulus set. Subjects were instructed to listen to the stimulus vigilantly without any explicit task in the passive-listening experiment. **b.** *Cocktail-party experiment.* A pair of stories told by individuals of different genders were selected from the single-speaker stimulus set and overlaid temporally to generate a two-speaker stimulus set. Subjects were instructed to attend either to the male or female speaker in the cocktail-party experiment. The same two-speaker story was presented twice in separate runs while the target speaker was varied. Attention condition was fixed within runs and it alternated across runs.

3.2 Materials and Methods

3.2.1 Participants

Functional data were collected from five healthy adult native subjects (four males and one female; aged between 26 and 31) who had no reported hearing problems and were native English speakers. The experimental procedures were approved by the Committee for the Protection of Human Subjects at University of California, Berkeley. Written informed consent was obtained from all subjects.

3.2.2 Stimuli

Figure 3.1 illustrates the two main types of stimuli used in the experiments: single-speaker stories and two-speaker stories. Ten single-speaker stories were

taken from The Moth Radio Program: “Alternate Ithaca Tom” by Tom Weiser; “How to Draw a Nekkid Man” by Tricia Rose Burt; “Life Flight” by Kimberly Reed; “My Avatar and Me” by Laura Albert; “My First Day at the Yankees” by Matthew McGough; “My Unhurried Legacy” by Kyp Malone; “Naked” by Catherine Burns; “Ode to Stepfather” by Ethan Hawke; “Targeted” by Jen Lee and “Under the Influence” by Jeffery Rudell. All stories were told before a live audience by a male or female speaker, and they were about 10-15 min long. Each two-speaker story was generated by temporally overlaying a pair of stories told by different genders and selected from the single-speaker story set. When the durations of the two single-speaker stories differed, the longer story was clipped from the end to match durations. Three two-speaker stories were prepared: from “Targeted” and “Ode to Stepfather” (cocktail1); from “How to Draw a Nekkid Man” and “My First Day at the Yankees” (cocktail2); and from “Life Flight” and “Under the Influence” (cocktail3). In the end, the stimuli consisted of ten single-speaker and three two-speaker stories.

3.2.3 Experimental procedures

Figure 3.1 outlines the two main experiments conducted in separate sessions: passive-listening and cocktail-party experiments. In the passive-listening experiment, subjects were instructed to listen to single-speaker stories vigilantly without an explicit attentional target. To facilitate sustained vigilance, we picked engaging spoken narratives from the Moth Radio Hour (see Stimuli). Each of the ten single-speaker stories was presented once in a separate run of the experiment. Two two-hour sessions were conducted, resulting in ten runs of passive-listening data for each subject. In the cocktail-party experiment, subjects were instructed to listen to two-speaker stories while attending to a target speaker (either the male or the female speaker). Our experimental design focuses on attentional modulations of speech representations when a stimulus is attended versus unattended. Each of the three cocktail-stories was presented twice in separate runs. This allowed us to present the same stimulus set in attended and unattended conditions to prevent potential biases due to across-condition stimulation differences. To

minimize adaptation effects, different two-speaker stories were presented in consecutive runs while maximizing the time window between repeated presentations of a two-speaker story. Attention condition alternated across consecutive runs. An exemplary sequence of runs was: cocktail1-M (attend to male speaker in cocktail1), cocktail2-F (attend to female speaker in cocktail2), cocktail3-M, cocktail1-F, cocktail2-M, and cocktail3-F. The first attention condition assigned to each two-speaker story (M or F) was counterbalanced across subjects. This resulted in a balanced assignment of ‘attended’ versus ‘unattended’ conditions during the second exposure to each two-speaker story. Furthermore, for each subject, the second exposure to half of the single-speaker stories (3 out of 6 included within the two-speaker stories) coincided with the ‘attended’ condition, whereas the second exposure to the other half coincided with the ‘unattended’ condition. Hence, second exposure to each story was balanced across ‘attended’ and ‘unattended’ conditions both within and across subjects. A two-hour session was conducted, resulting in six runs of cocktail-data. Note that the two-speaker stories used in the cocktail-party experiment were constructed from the single-speaker story set used in passive-listening experiment. Hence, for each subject, the cocktail-party experiment was conducted several months (5.5 months) after the completion of the passive-listening experiment to minimize potential repetition effects. The dataset collected from the passive-listening experiment was previously analyzed [143, 119]; however, the dataset collected from the cocktail-party experiment was specifically collected for this study.

In both experiments, the length of each run was tailored to the length of the story stimulus with additional 10 sec of silence both before and after the stimulus. All stimuli were played at 44.1 kHz and delivered binaurally to both ears using Sensimetrics S14 in-ear piezo-electric headphones. The Sensimetrics S14 is an MRI-compatible auditory stimulation system with foam canal tips to reduce scanner noise (above 29 dB as stated in specifications). The frequency response of the headphones was flattened using a Behringer Ultra-Curve Pro Parametric Equalizer. Furthermore, the level of sound was adjusted for each subject to ensure clear and comfortable hearing of the stories.

3.2.4 MRI data collection and preprocessing

MRI data were collected on a 3T Siemens TIM Trio scanner at the Brain Imaging Center, UC Berkeley, using a 32-channel volume coil. For functional scans, a gradient echo EPI sequence was used with $TR = 2.0045$ s, $TE = 31$ ms, flip angle $= 70^\circ$, voxel size $= 2.24 \times 2.24 \times 4.1$ mm³, matrix size $= 100 \times 100$, field of view $= 224 \times 224$ mm² and 32 axial slices covering the entire cortex. For anatomical data, a T1-weighted multi-echo MP-RAGE sequence was used with voxel size $= 1 \times 1 \times 1$ mm³ and field of view $= 256 \times 212 \times 256$ mm³.

Each functional run was motion corrected using FMRIB’s Linear Image Registration Tool (FLIRT) [144]. A cascaded motion-correction procedure was performed, where separate transformation matrices were estimated within single runs, within single sessions and across sessions sequentially. To do this, volumes in each run were realigned to the mean volume of the run. For each session, the mean volume of each run was then realigned to the mean volume of the first run in the session (see Supp. Table A.1 for within-session motion statistics during the cocktail-party experiment). Lastly, the mean volume of the first run of each session was realigned to the mean volume of the first run of the first session of the passive-listening experiment. The estimated transformation matrices were concatenated and applied in a single step. Motion-corrected data were manually checked to ensure that no major realignment errors remained. The moment-to-moment variations in head position were also estimated and used as nuisance regressors during model estimation to regress out motion-related nuisance effects from BOLD responses. The Brain Extraction Tool (BET) in FSL 5.0 [145] was used to remove non-brain tissues. This resulted in 68016-84852 brain voxels in individual subjects. All model fits and analyses were performed on these brain voxels in volumetric space.

3.2.5 Visualization on cortical flatmaps

Cortical flatmaps were used for visualization purposes, where results in volumetric space were projected onto the cortical surfaces using PyCortex [146]. Cortical surfaces were reconstructed from anatomical data using Freesurfer [147]. Five relaxation cuts were made into the surface of each hemisphere, and the surface crossing the corpus callosum was removed. Functional data were aligned to the individual anatomical data with affine transformations using FLIRT [144]. Cortical flatmaps were constructed for visualization of significant model prediction scores, functional selectivity and attentional modulation profiles, and representational complexity and modulation gradients.

3.2.6 ROI definitions and abbreviations

We defined region of interests for each subject based on an atlas-based parcellation of the cortex [148]. To do this, functional data were co-registered to the individual-subject anatomical scans with affine transformations using FLIRT [144]. Individual-subject anatomical data were then registered to the Freesurfer standard anatomical space via the boundary-based registration tool in FSL [149]. This procedure resulted in subject-specific transformations mapping between the standard anatomical space and the functional space of individual subjects. Anatomical regions of interest from the Destrieux atlas were outlined in the Freesurfer standard anatomical space; and they were back-projected onto individual-subject functional spaces via the subject-specific transformations using PyCortex [146]. The anatomical regions were labeled according to the atlas. To explore potential selectivity gradients across the lateral aspects of Superior Temporal Gyrus and Superior Temporal Sulcus, these ROIs were further split into three equidistant sub-regions in posterior-to-anterior direction. Heschl's Gyrus and Heschl's Sulcus were considered as a single ROI as prior reports suggest that primary auditory cortex is not constrained by Heschl's Gyrus and extends to Heschl's Sulcus as well [103, 104, 150]. We only considered regions with at least ten speech-selective voxels in each individual subject for subsequent analyses.

Supplementary Table A.2 and A.3 lists the defined ROIs and the number of spectrally, articulatorily and semantically selective voxels within each ROI in the left and right hemisphere, with number of speech-selective voxels. ROI abbreviations and corresponding Destrieux indices are: Heschl’s Gyrus and Heschl’s Sulcus (HG/HS: 33 and 74), Planum Temporale (PT: 36), posterior segment of Sylvian Fissure (pSF: 41), lateral aspect of Superior Temporal Gyrus (STG: 34), Superior Temporal Sulcus (STS, 73), Middle Temporal Gyrus (MTG: 38), Angular Gyrus (AG: 25), Supramarginal Gyrus (SMG: 26), Intraparietal Sulcus (IPS: 56), opercular part of Inferior Frontal Gyrus/Pars Opercularis (POP: 12), triangular part of Inferior Frontal Gyrus/Pars Triangularis (PTR: 14), Precentral Gyrus (PreG: 29), medial Occipito-Temporal Sulcus (mOTS:60), Inferior Frontal Sulcus (IFS: 52), Middle Frontal Gyrus (MFG:15), Middle Frontal Sulcus (MFS: 53), Superior Frontal Sulcus (SFS: 54), Superior Frontal Gyrus (SFG: 16), Precuneus (PreC: 30), Subparietal Sulcus (SPS: 71), and Posterior Cingulate Cortex (PCC: 9 and 10). The subregions of STG are: aSTG (anterior one third of STG), mSTG (middle one third of STG) and pSTG (posterior one third of STG). The subregions of STS are: aSTS (anterior one third of STS), mSTS (middle one third of STS) and pSTS (posterior one third of STS). MTG was not split into subregions since these subregions did not have a sufficient number of speech-selective voxels in each individual subject.

3.2.7 Model construction

To comprehensively assess speech representations, we constructed spectral, articulatory, and semantic models of the speech stimuli (Figure 3.2; [119]).

3.2.7.1 Spectral model

For the spectral model, cochleogram features of speech were estimated based on Lyon’s Passive Ear model. Lyon’s human cochlear model involves logarithmic filtering, compression and adaptive gain control operations applied to input

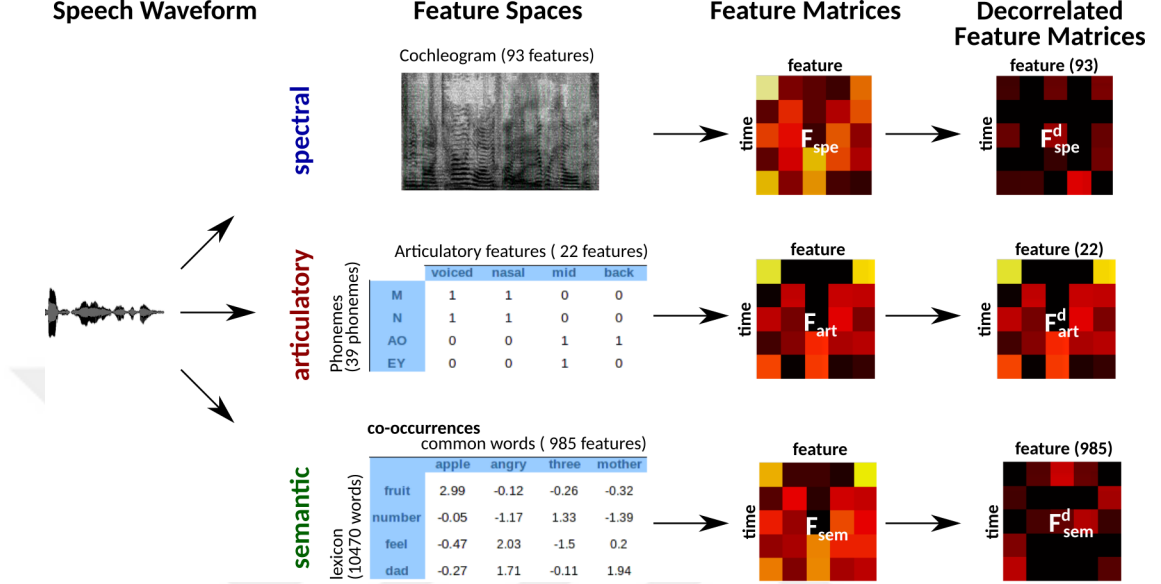


Figure 3.2: **Multi-level speech features.** Three distinct feature spaces were constructed to represent natural speech at multiple levels: spectral, articulatory, and semantic spaces. Speech waveforms were projected separately on these spaces to form stimulus matrices. The spectral feature matrix captured the cochleogram features of the stimulus in 93 channels having center frequencies between 115 and 9920 Hz. The articulatory feature matrix captured the mapping of each phoneme in the stimulus to 22 binary articulation features. The semantic feature matrix captured the statistical co-occurrences of each word in the stimulus with 985 common words in English. Each feature matrix was Lanczos-filtered at a cutoff frequency of 0.25 Hz and downsampled to 0.5 Hz to match the sampling rate of fMRI. Natural speech might contain intrinsic stimulus correlations among spectral, articulatory, and semantic features. To prevent potential biases due to stimulus correlations, we decorrelated the three feature matrices examined here via Gram-Schmidt orthogonalization (see Methods). The decorrelated feature matrices were used for modeling BOLD responses.

sound [151, 152, 153]. Depending on the sampling rate of the input signal, the cochlear model generates 118 waveforms with center frequencies between 84 Hz and 21 kHz. Considering the frequency response of the headphones used in the experiment, 93 waveforms with center frequencies between 115 Hz and 9920 Hz were selected as the features of the spectral model. The spectral features were Lanczos-filtered at a cutoff frequency of 0.25 Hz and downsampled to 0.5 Hz to match the sampling rate of fMRI. The 93 spectral features were then temporally z-scored to zero mean and unit variance.

3.2.7.2 Articulatory model

For the articulatory model, each phoneme in the stories was mapped onto a unique set of 22 articulation features; for example, phoneme /ZH/ is postalveolar, fricative and voiced (Table 3.1; [154, 119]). This mapping resulted in 22-dimensional binary vectors for each phoneme. To obtain the timestamp of each phoneme and word in the stimuli, the speech in the stories were aligned with the story transcriptions using the Penn Phonetics Lab Forced Aligner [155]. Alignments were manually verified and corrected using Praat (www.praat.org). The articulatory features were Lanczos-filtered at a cutoff frequency of 0.25 Hz and downsampled to 0.5 Hz. Finally, the 22 articulatory features were z-scored to zero mean and unit variance.

3.2.7.3 Semantic model

For the semantic model, co-occurrence statistics of words were measured via a large corpus of text [156, 143, 119]. The text corpus was compiled from 2,405,569 Wikipedia pages, 36,333,459 user comments scraped from reddit.com, 604 popular books and the transcripts of 13 Moth stories (including the stories used as stimuli). We then built a 10,470-word lexicon from the union set of the 10,000 most common words in the compiled corpus and all words appearing in the ten Moth stories used in the experiment. Basis words were then selected as a set of 985 unique words from Wikipedia’s List of 1000 Basic Words. Co-occurrence statistics of the lexicon words with 985 basis words within a 15-word window were characterized as a co-occurrence matrix of size 985x10,470. Elements of the resulting co-occurrence matrix were log-transformed, z-scored across columns to correct for differences in basis-word frequency, and z-scored across rows to correct for differences in lexicon-word frequency. Each word in the stimuli was then represented with a 985-dimensional co-occurrence vector based on the speech-transcription alignments. The semantic features were Lanczos-filtered at a cutoff frequency of 0.25 Hz and downsampled to 0.5 Hz. The 985 semantic features were finally z-scored to zero mean and unit variance.

Table 3.1: Phoneme-articulatory features map

Phonemes	Articulatory Features			
/B/	Bilabial	Plosive	Voiced	
/CH/	Postalveolar	Affricate	Unvoiced	
/D/	Alveolar	Plosive	Voiced	
/DH/	Dental	Fricative	Voiced	
/F/	Labiodental	Fricative	Unvoiced	
/G/	Velar	Plosive	Voiced	
/HH/	Glottal	Fricative	Unvoiced	
/JH/	Postalveolar	Affricate	Voiced	
/K/	Velar	Plosive	Unvoiced	
/L/	Alveolar	Lateral	Voiced	
/M/	Bilabial	Nasal	Voiced	
/N/	Alveolar	Nasal	Voiced	
/NG/	Velar	Nasal	Voiced	
/P/	Bilabial	Plosive	Unvoiced	
/R/	Alveolar	Approximant	Voiced	
/S/	Alveolar	Fricative	Unvoiced	
/SH/	Postalveolar	Fricative	Unvoiced	
/T/	Alveolar	Plosive	Unvoiced	
/TH/	Dental	Fricative	Unvoiced	
/V/	Labiodental	Fricative	Voiced	
/W/	Velar	Approximant	Voiced	
/Y/	Palatal	Approximant	Voiced	
/Z/	Alveolar	Fricative	Voiced	
/ZH/	Postalveolar	Fricative	Voiced	
/AA/	Low	Back		
/AE/	Low	Front		
/AH/	Mid	Central		
/AO/	Mid	Back		
/AW/	Low	Central	Mid	Back
/AY/	Low	Central	Mid	Front
/EH/	Mid	Front		
/ER/	Mid	Central		
/EY/	Mid	Front		
/IH/	Mid	Front		
/IY/	High	Front		
/OW/	Mid	Back		
/OY/	Mid	Back	High	Front
/UH/	High	Back		
/UW/	High	Back		

3.2.7.4 Decorrelation of feature spaces.

In natural stories, there might be potential correlations among certain spectral, articulatory, or semantic features. If significant, such correlations can partly confound assessments of model performance. To assess the unique contribution of each feature space to the explained variance in BOLD responses, a decorrelation procedure was first performed (Figure 3.2). To decorrelate a feature matrix F of size $m \times n$ from a second feature matrix K of size $m \times p$, we first found an orthonormal basis for the column space of K ($col\{K\}$) using economy-size singular value decomposition:

$$K_{m \times p} = U_{m \times p} \times S_{p \times p} \times V_{p \times p} \quad (3.1)$$

where U contains left singular vectors as columns, V contains right singular vectors, and S contains the singular values. Left singular vectors were taken as the orthonormal basis for $col\{K\}$, and each column of F was decorrelated from it according to the following formula:

$$\vec{f}_i^d = \vec{f}_i - \sum_{j=1}^p \left(\vec{f}_i \cdot \vec{u}_j \right) \cdot \vec{u}_j \quad (3.2)$$

Where $\vec{f}_i, \vec{u}_j$ are the column vectors of F and U respectively, and $\vec{f}_i^d$ is the column vectors of the decorrelated feature matrix, F^d . To decorrelate feature matrices for the models considered here, we took the original articulatory feature matrix as a reference, and decorrelated the spectral feature matrix from the articulatory feature matrix, and decorrelated the semantic feature matrix from both articulatory and spectral feature matrices. This decorrelation sequence was selected because spectral and articulatory features capture lower-level speech representations, and the articulatory feature matrix had the fewest number of features among all models. In the end, we obtained 3 decorrelated feature matrices whose columns had zero correlation with the columns of the other two matrices.

3.2.8 Analyses

The main motivation of this study is to understand whether and how strongly various levels of speech representations are modulated across cortex during a cocktail-party task. To answer this question, we followed a two-stage approach as illustrated in Figure 3.3. In the first stage, we identified voxels selective for speech features using data from the passive-listening experiment. To do this, we measured voxelwise selectivity separately for spectral, articulatory, and semantic features of the single-speaker stories. In the second stage, we used the models fit using passive-listening data to predict BOLD responses measured in the cocktail-party experiment. Prediction scores for attended versus unattended stories were compared to quantify the degree of attentional modulations, separately for each model and globally across all models.

Note that a subset of the ten single-speaker-stories were used to generate three two-speaker-stories used in the experiments. To prevent potential bias, a three-fold cross-validation procedure was performed for testing models fit using passive-listening data on cocktail-party data. In each fold, models were fit using eight-run passive-listening data; and separately tested on two-run passive-listening data and two-run cocktail-party data. The same set of test stories were used both in the passive-listening and cocktail-party experiments to minimize risk of poor model generalization between the passive-listening and cocktail-party experiments due to uncontrolled stimulus differences. There was no overlap between the stories in the training and testing runs. Model predictions were aggregated across three folds, and prediction scores were then computed.

3.2.8.1 Voxelwise modeling

In the first stage, we fit voxelwise models in individual subjects using passive-listening data. To account for hemodynamic delays, we used a linearized four-tap finite impulse response (FIR) filter to allow different HRF shapes for separate brain regions (Goutte et al. 2000). Each model feature was represented as four

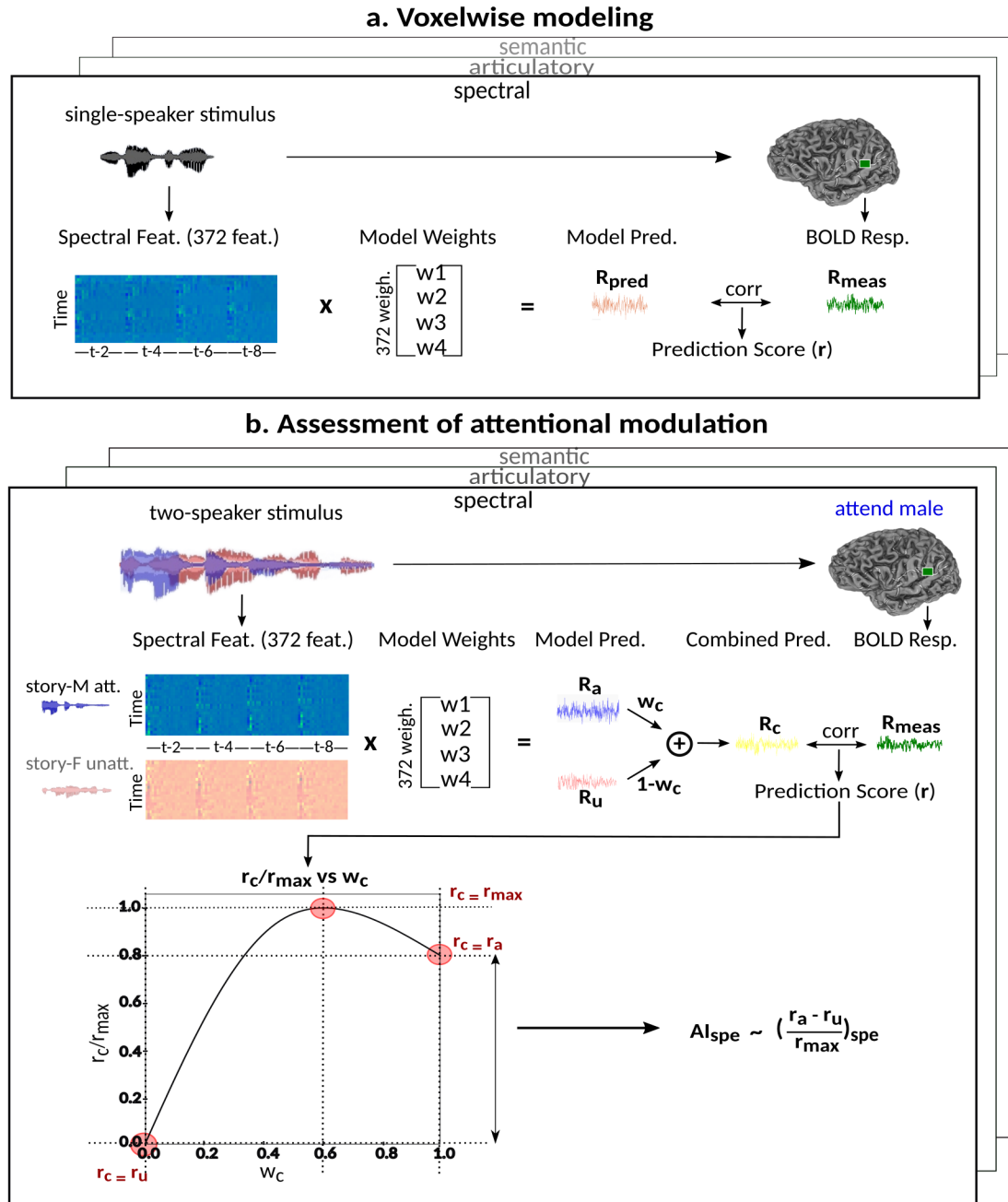


Figure 3.3: Caption next page

Figure 3.3: **Modeling procedures.** **a. Voxelwise modeling.** Voxelwise models were fit in individual subjects using passive-listening data. To account for hemodynamic response, a linearized four-tap finite impulse response (FIR) filter spanning delayed effects at 2-8 sec was used. Models were fit via L2-regularized linear regression. BOLD responses were predicted based on fit voxelwise models on held-out passive-listening data. Prediction scores were taken as the Pearson’s correlation between predicted and measured BOLD responses. For a given subject, speech-selective voxels were taken as the union of voxels significantly predicted by spectral, articulatory, or semantic models ($q(FDR) < 10^{-5}, t\text{-test}$). **b. Assessment of attentional modulation.** Passive-listening models for single voxels were tested on cocktail-party data to quantify attentional modulations in selectivity. In a given run, one of the speakers in a two-speaker story was attended while the other speaker was ignored. Separate response predictions were obtained using the isolated story stimuli for the attended speaker and for the unattended speaker. Since a voxel can represent information from both attended and unattended stimuli, a linear combination of these predicted responses was considered with varying combination weights (wc in $[0\ 1]$). BOLD responses were predicted based on each combination weight separately. Three separate prediction scores were calculated based on only the attended stimulus ($wc = 1$), based on only the unattended stimulus ($wc = 0$), and based on the optimal combination of the two stimuli. A model-specific attention index, (AI_m) was then computed as the ratio of the difference in prediction scores for attended versus unattended stories to the prediction score for their optimal combination (see Methods).

features in the stimulus matrix to account for their delayed effects in BOLD responses at 2, 4, 6 and 8 sec. Model weights, W , were then found using L2-regularized linear regression:

$$W = (F^T F + I\lambda)^{-1} F^T R \quad (3.3)$$

Here, λ is the regularization parameter, F is the decorrelated feature matrix for a given model and R is the aggregate BOLD response matrix for cortical voxels. A cross-validation procedure with 50 iterations was performed to find the best regularization parameter for each voxel among 30 equispaced values in log-space of $1 : 10^5$. The training passive-listening data was split into 50 equisized chunks, where 1 chunk was reserved for validation and 49 chunks were reserved for model fitting at each iteration. Prediction scores were taken as Pearson’s correlation between predicted and measured BOLD responses. The optimal λ value for each voxel was selected by maximizing the average prediction score

across cross-validation folds. The final model weights were obtained using the entire set of training passive-listening data and the optimal λ . Next, we measured the prediction scores of the fit models on testing data from the passive-listening experiment. Spectrally-, articulatorily-, and semantically-selective voxels were separately identified in each ROI based on the set of significantly-predicted voxels by each model. A given ROI was considered selective for a model, only if it contained ten or more significant voxels for that model ($q(FDR) < 10^{-5}, t\text{-test}$). Speech-selective voxels within the ROI were then taken as the union of these spectrally-, articulatorily-, and semantically-selective voxels. Subsequent analyses were performed on speech-selective voxels.

Model-specific selectivity index. Single-voxel prediction scores on passive-listening data were used to quantify the degree of selectivity of each ROI to the underlying model features under passive-listening. To do this, a model-specific selectivity index, (SI_m) , was defined as follows:

$$SI_m = \frac{(r)_m}{\sum_i (r)_i}, \quad i, m \in \{spe, art, sem\} \quad (3.4)$$

where r is the average prediction score across speech-selective voxels within the ROI during passive-listening. SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model.

Complexity index. The complexity of speech representations was characterized via a complexity index, (CI) , which reflected the relative tuning of an ROI for low- versus high-level speech features. The following intrinsic complexity levels were assumed for the three speech models considered here: $(c_{spe}, c_{art}, c_{sem}) = (0.0, 0.5, 1.0)$. Afterwards, CI was taken as the average of the complexity levels weighted by the selectivity indices:

$$CI = \sum_m SI_m c_m, \quad m \in \{spe, art, sem\} \quad (3.5)$$

CI is in the range of $[0, 1]$, where higher values indicate stronger tuning for

semantic features and lower values indicate stronger tuning for spectral features.

3.2.8.2 Assessment of attentional modulations

In the second stage, we tested the passive-listening models on cocktail-party data to quantify ROI-wise attentional modulation in selectivity for corresponding model features and to find the extent of the representation of unattended speech. These analyses were repeated separately for the three speech models.

Model-specific attention index. To quantify the attentional modulation in selectivity for speech features, we compared prediction scores for attended versus unattended stories in the cocktail-party experiment. Models fit using passive-listening data were used to predict BOLD responses elicited by two-speaker stories. In each run, only one of the speakers in a two-speaker story was attended while the other speaker was ignored. Separate response predictions were obtained using the isolated story stimuli for the attended and unattended speakers. Since a voxel can represent information on both attended and unattended stimuli, a weighted linear combination of these predicted responses was considered:

$$R_c = R_a w_c + R_u (1 - w_c) \quad (3.6)$$

where R_a and R_u are the predicted responses for the attended and unattended stories in a given run; R_c is the combined response and w_c is the combination weight. We computed R_c for each separate w_c value in $[0:0.1:1]$. Note that $R_c = R_a$ when $w_c = 1.0$; and $R_c = R_u$ when $w_c = 0.0$. We then calculated single-voxel prediction scores for each w_c value. An illustrative plot of r_c/r_{max} is given in Figure 3.3b, where r_c denotes the prediction scores and r_{max} denotes the maximum r_c value (the optimal combination). r_a and r_u are the prediction scores for attended and unattended stories respectively. To quantify the degree of attentional modulation, a model-specific attention index (AI_m) was taken as:

$$AI_m = \alpha_m \left(\frac{r_a - r_u}{r_{max}} \right)_m, \quad \alpha_m = \frac{(r_{max})_m}{\sum_i (r_{max})_i}, \quad m, i \in \{spe, art, sem\}, \quad (3.7)$$

where r_{max} denotes an ideal upper limit for model performance, and α_m reflects the relative model performance under the cocktail-party task. Note that AI_m considers selectivity to the underlying model features when calculating the degree of attentional modulation.

Global attention index. We then computed global attention index (gAI) as follows:

$$gAI = \sum_m AI_m, \quad m \in \{spe, art, sem\} \quad (3.8)$$

Both gAI and AI_m are in the range $[-1, 1]$. A positive index indicates attentional modulation of selectivity in favor of the attended stimuli and a negative index indicates attentional modulation in favor of the unattended stimuli. A value of zero indicates no modulation.

3.2.8.3 Colormap in selectivity and modulation profile flatmaps.

The cortical flatmaps of selectivity and modulation profiles use a colormap that shows the relative contributions of all three models to the selectivity and attention profiles. For selectivity profiles, a continuous colormap was created by assigning significantly positive articulatory, semantic and spectral selectivity to the red, green and blue (R, G, B) color channels, respectively. During assignment, selectivity values were normalized to sum of one, and then normalized to linearly map the interval $[0.15 \ 0.85]$ to $[0 \ 1]$. Distinct colors were assigned to six landmark selectivity values: red for $(1, 0, 0)$, green for $(0, 1, 0)$, blue for $(0, 0, 1)$, yellow for $(0.5, 0.5, 0)$, magenta for $(0.5, 0, 0.5)$, and turquoise for $(0, 0.5, 0.5)$. The same procedures were also applied for creating a colormap for modulation profiles.

3.2.9 Statistical tests

3.2.9.1 Significance assessments within subjects

For each voxel-wise model, significance of prediction scores was assessed via a t-test; and resulting p-values were false-discovery-rate corrected for multiple comparisons (FDR; Benjamini and Hochberg 1995).

A bootstrap test was used in assessments of SI_m , CI , AI_m and gAI within single subjects. In ROI analyses, speech-selective voxels within a given ROI were resampled with replacement 10000 times. For each bootstrap sample, mean prediction score of a given model was computed across resampled voxels. Significance level was taken as the fraction of bootstrap samples in which the test metric computed from these prediction scores is less than 0 (for right-sided tests) or greater than 0 (for left-sided tests). The same procedure was also used for comparing pairs of ROIs, where ROI voxels were resampled independently.

3.2.9.2 Significance assessments across subjects

A bootstrap test was used in assessments of SI_m , CI , AI_m and gAI across subjects. In ROI analyses, ROI-wise metrics were resampled across subjects with replacement 10000 times. Significance level was taken as the fraction of bootstrap samples where the test metric averaged across resampled subjects is less than 0 (for right-sided tests) or greater than 0 (for left-sided tests). The same procedure was also used for comparisons among pairs of ROIs. Here, we used a more stringent significance definition for across-subjects tests that focuses on effects consistently observed in each individual subject. Therefore, an effect was taken significant only if the same metric was found significant in each individual subject.

3.3 Results

3.3.1 Intrinsic cortical representations of natural speech

To examine the cortical distribution and strength of attentional modulations in speech representations, we first obtained a baseline measure of intrinsic selectivity for speech features. For this purpose, we fit voxelwise models using BOLD responses recorded during passive listening. Speech representations are thought to be organized hierarchically across multiple stages of processing in the brain, ranging from acoustic features in early auditory cortex to linguistic features in downstream areas [115, 18, 117, 157, 158]. To broadly examine this hierarchy, we built three separate models containing low-level spectral, intermediate-level articulatory and high-level semantic features of natural stories [119]. Figure 3.4 displays the cortical distribution of prediction scores for each model in a representative subject, and Supp. Table A.2 and A.3 list the number of significantly predicted voxels by each model in anatomical ROIs. We find spectrally-selective voxels mainly in early auditory regions (bilateral HG/HS and PT; and left pSF) and bilateral SMG, and articulatorily-selective voxels mainly in early auditory regions (bilateral HG/HS and PT; and left pSF), bilateral STG, STS, SMG and MFS as well as left POP and PreG. In contrast, semantically-selective voxels are found broadly across cortex except early auditory regions (bilateral HG/HS and right PT).

To quantitatively examine cortical overlap among spectral, articulatory, and semantic representations, we separately measured the degree of functional selectivity for each feature level via a model-specific selectivity index (SI_m ; see Methods). Bar plots of selectivity indices are displayed in Figure 3.5a for perisylvian cortex and in Figure 3.6 for non-perisylvian cortex (see Supp. Figs. A.1-5 for single-subject results). Distinct selectivity profiles are observed from distributed selectivity for spectral, articulatory, and semantic features (e.g., left PT and right pSTG) to strong tuning to a single level of features (e.g., left IPS and PCC). The selectivity profiles of the ROIs are also visualized on the cortical flatmap projecting articulatory, semantic, and spectral selectivity indices of each ROI to the

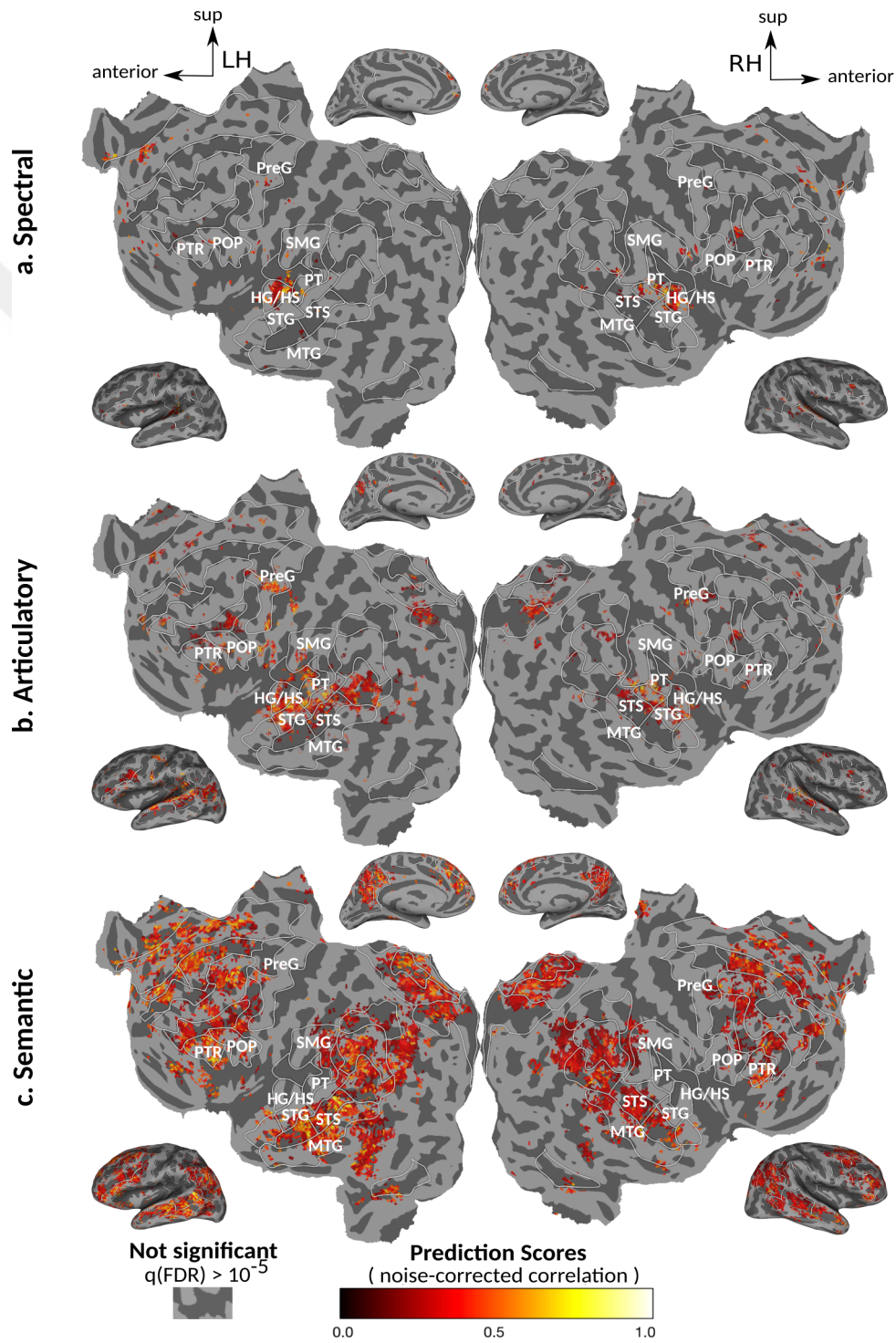


Figure 3.4: Caption next page

Figure 3.4: **Prediction scores of voxelwise speech models.** **a.** *Spectral model.* Prediction scores for voxels significantly predicted by the spectral model are plotted on the flattened cortical surface of subject S4. Medial and lateral views of the inflated hemispheres are also shown above and below the flatmaps, respectively. Colors indicate the value of the noise-corrected prediction scores (see legend). White lines encircle ROIs that are found based on an automatic atlas-based cortical parcellation. Labels of some relevant ROIs are shown (see Methods for ROI abbreviations). In this subject, significantly predicted voxels by the spectral model lie mainly in the dorsal aspect of the superior temporal cortex. **b.** *Articulatory model.* Significantly predicted voxels by the articulatory model are located mainly in the dorsal and lateral aspects of the superior temporal cortex extending to STS, and in inferior frontal regions. **c.** *Semantic model.* Significantly predicted voxels by the semantic model are broadly distributed across cortex covering most of the temporal, prefrontal and parietal cortical regions except the early auditory cortex.

red, green, and blue channels of the RGB colormap as seen in Figure 3.5b (see Supp. Fig. A.6 for selectivity profile flatmaps in individual subjects; see Methods for colormap details). A progression from low-intermediate to high-level speech representations is apparent across bilateral temporal cortex in superior-inferior direction ($HG/HS \rightarrow mSTG \rightarrow mSTS \rightarrow MTG$) consistently in all subjects. Furthermore, many higher-order regions in parietal (bilateral AG, IPS, SPS, PrC, PCC and POS) and frontal cortices (bilateral PTR, IFS, MFG, SFS, and SFG; and left POP) manifest dominant semantic selectivity consistently in all subjects ($p < 0.05$; see Supp. Figs. A.1-5 for single-subject results). These results corroborate the view that speech representations are hierarchically organized across cortex with partial overlap mostly in early and intermediate stages of speech processing.

3.3.1.1 Hierarchies in speech representations

An emerging view is that speech features represented in the cortex grow progressively more complex towards downstream areas [159, 18, 19, 157, 119]. Yet, a broad quantitative characterization of these selectivity gradients is lacking. To assess the selectivity in each stage of processing, we first measured a complexity index, CI , that reflects whether an ROI is relatively tuned for low-level spectral or high-level semantic features (see Methods). CI across perisylvian and

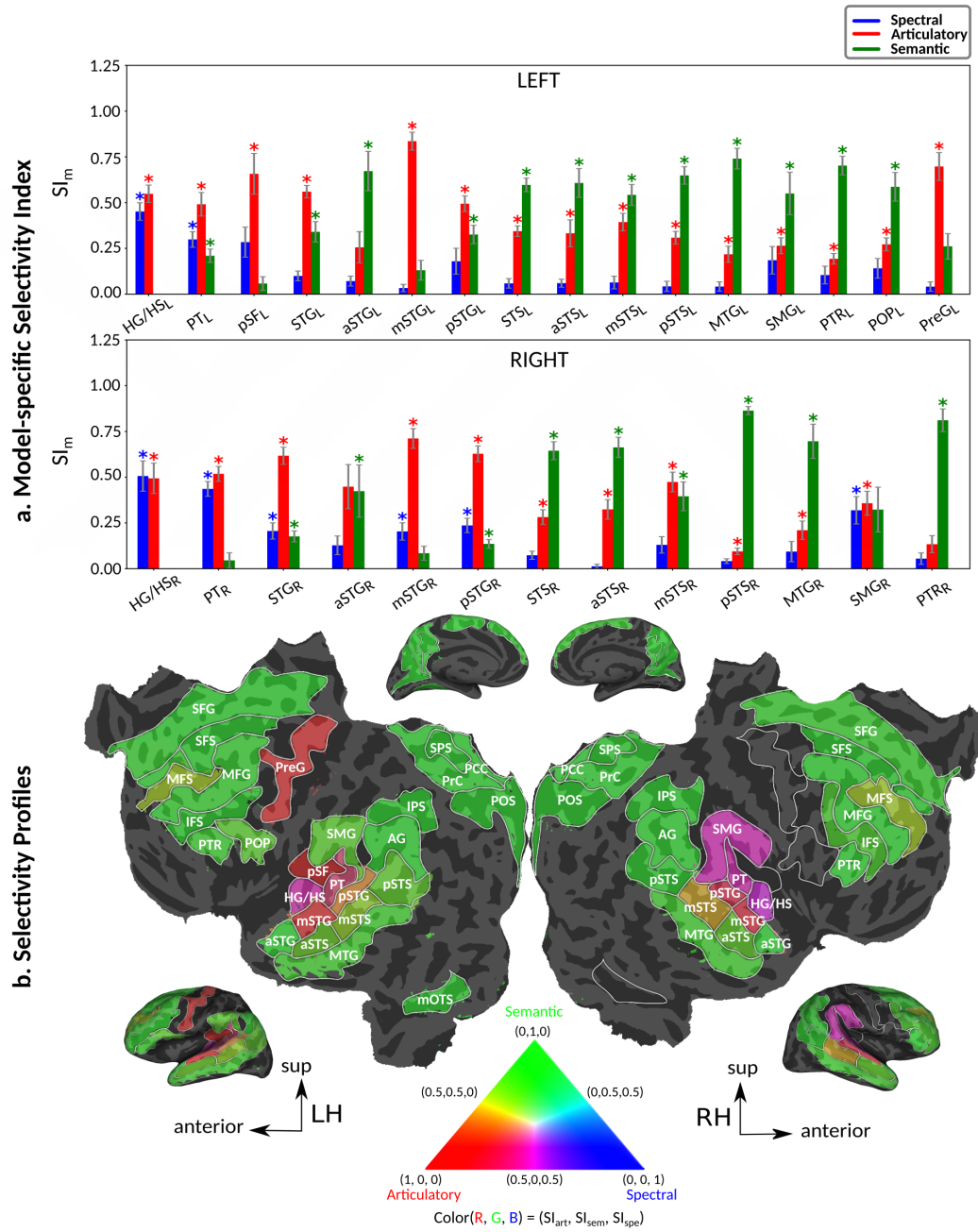


Figure 3.5: Caption next page

Figure 3.5: **Selectivity for multi-level speech features.** **a.** *Model-specific selectivity indices.* Single-voxel prediction scores on passive-listening data were used to quantify the selectivity of each ROI to underlying model features. Model-specific prediction scores were averaged across speech-selective voxels within each ROI and normalized such that the cumulative score from all models was 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory, and semantic models (mean $\pm$ sem across subjects). Significant indices are marked with * ($p < 0.05$; see Sup. Figs. A.1-5 for selectivity indices of individual subjects). ROIs in perisylvian cortex are displayed (see Figure 3.6 for non-perisylvian ROIs; see Methods for ROI abbreviations). ROIs in LH and RH are shown in the top and bottom panels, respectively. POP_R and PreG_R that did not have consistent speech selectivity in individual subjects were excluded (see Methods). **b.** *Intrinsic selectivity profiles.* Selectivity profiles of cortical ROIs averaged across subjects are shown on the cortical flatmap of a representative subject (S4). Significant articulatory, semantic, and spectral selectivity indices of each ROI are projected to the red, green, and blue channels of the RGB colormap (see Methods). This analysis only included ROIs with consistent selectivity for speech features in each individual subject. Medial and lateral views of the inflated hemispheres are also shown. A progression from low-intermediate to high-level speech representations are apparent across bilateral temporal cortex in the superior-inferior direction; consistently in all subjects (see Supp. Fig. A.6 for selectivity profiles of individual subjects). Meanwhile, semantic selectivity is dominant in many higher-order regions within the parietal and frontal cortices (bilateral AG, IPS, SPS, PrC, PCC, POS, PTR, IFS, SFS, SFG, MFG and left POP) ($p < 0.05$; see Supp. Figs. A.1-5). These results support the view that speech representations are hierarchically organized across cortex with partial overlap between spectral, articulatory and semantic representations in early to intermediate stages of auditory processing..

non-perisylvian cortex is displayed in Figure 3.7. Early auditory regions such as HG/HS have CI close to 0, whereas higher-order regions like PTR have CI tending to 1 as expected.

Next, we explored in detail finer scale gradients across the main auditory streams: dorsal and ventral stream [160, 17, 18, 159, 161, 19, 162, 163].

Dorsal stream. The dorsal stream emanates from primary auditory cortex (PAC) [164, 163, 165], traverses planum temporale [17, 18, 19], and then projects to pars-operculum/BA44 [20, 16, 166] and precentral-gyrus/BA6 [167, 20, 16] in prefrontal cortex. The projections onto prefrontal cortex are suggested to be

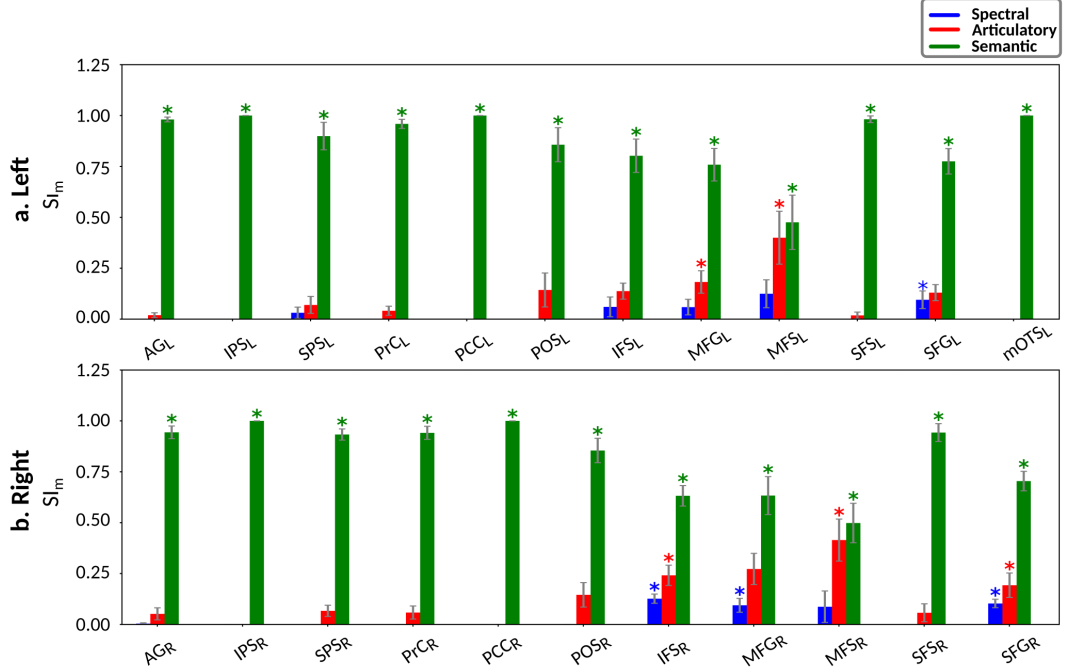


Figure 3.6: **Model-specific selectivity indices for non-perisylvian ROIs.** Single-voxel prediction scores on passive-listening data were used to quantify the degree of selectivity of each ROI to the underlying model features under passive listening. Prediction scores for a given model were averaged across speech-selective voxels within each ROI, and then normalized such that the cumulative prediction score from all models summed to 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory and semantic models (mean $\pm$ sem across subjects). Significant indices are marked with *, colored according to the model ($p < 0.05$, bootstrap test; see Supp. Figs. A.1-5 for selectivity indices of individual subjects). Only ROIs in non-perisylvian cortex are displayed. **a.** ROIs in LH. **b.** ROIs in RH. mOTSL that did not have consistent speech selectivity in individual subjects was excluded (see Methods). These results show that selectivity is primarily semantic in most of the non-perisylvian ROIs.

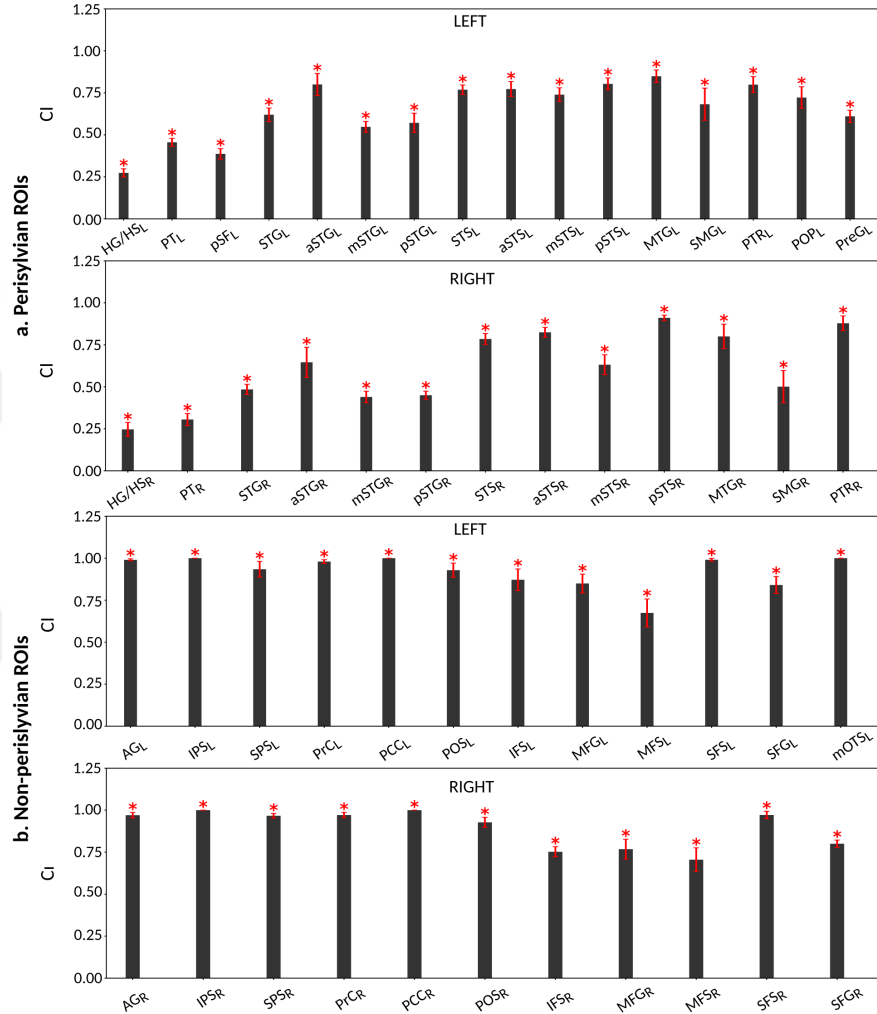


Figure 3.7: **Representational complexity.** To characterize the complexity of speech representations, a complexity index, (CI), was taken as the relative tuning of an ROI for low- versus high-level speech features. CI is in the range of $[0, 1]$, where higher values indicate stronger tuning for semantic features and lower values indicate stronger tuning for spectral features. Bar plots indicate CI (mean $\pm$ sem across subjects). Indices that are consistently significant in each individual subject are marked with * ($p < 10^{-4}$, bootstrap test). **a. Perisylvian ROIs.** ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. Early auditory regions such as HG/HS have CI close to 0, whereas higher-order regions like PTR have CI tending to 1. These results indicate that more complex features are represented in higher-order regions compared to the earlier ones. **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. Higher-order regions in non-perisylvian cortex mostly have similar representational complexity that is close to 1.

either direct or relayed via supramarginal-gyrus/BA40 in inferior parietal cortex [168, 12, 169, 170, 20]. In the light of these reports, we examined variation of CI across three trajectories as shown in Figure 3.8a: left dorsal-1 ($HG/HS_L \rightarrow PT_L \rightarrow (SMG_L) \rightarrow POP_L$), left dorsal-2 ($HG/HS_L \rightarrow PT_L \rightarrow (SMG_L) \rightarrow PreG_L$) and right dorsal ($HG/HS_R \rightarrow PTR \rightarrow SMG_R$). Left dorsal-1 and dorsal-2 were considered with or without inclusion of SMG. As shown in Figure 3.8b, we find significant increase in CI consistently in all subjects across the following left dorsal subtrajectories ($p < 0.05$): $CI_{HG/HS} < CI_{PT} < CI_{POP}$ and $CI_{HG/HS} < CI_{PT} < CI_{PreG}$. In contrast, we find no consistent pattern in CI across the right dorsal stream ($p > 0.05$). Note that PreG has greater articulatory selectivity in all subjects compared to POP ($p < 0.05$; see Supp. Figs. A.1-5). Furthermore, articulatory selectivity is dominant in PreG, while semantic selectivity is dominant in POP ($p < 0.05$; see Supp. Figs. A.1-5 and Supp. Fig. A6). These results indicate that left dorsal-1 and left dorsal-2 differ in that the former carries relatively stronger semantic representations whereas the latter hosts articulatory representations.

Ventral stream. The ventral stream emanates from PAC [164, 165], passes through middle-anterior superior temporal cortex [164, 157, 165], and then projects to pars-triangularis/BA45 [164, 166] in prefrontal cortex. It is also suggested to traverse superior-inferior paths in temporal lobe [17, 18]. As such, we examined variation of CI across four trajectories (Figure 3.8a): left ventral-1 ($HG/HS_L \rightarrow mSTG_L \rightarrow mSTS_L \rightarrow MTG_L$), left ventral-2 ($HG/HS_L \rightarrow mSTG_L \rightarrow aSTG_L \rightarrow PTR_L$), right ventral-1 ($HG/HS_R \rightarrow mSTG_R \rightarrow mSTS_R \rightarrow MTG_R$) and right ventral-2 ($HG/HS_R \rightarrow mSTG_R \rightarrow aSTG_R \rightarrow PTR_R$). As shown in Figure 3.8b, we find significant increase in CI consistently in all subjects across the following subtrajectories ($p < 0.05$): $CI_{HG/HS} < CI_{mSTG} < CI_{aSTG}$ and $CI_{HG/HS} < CI_{mSTG} < CI_{mSTS}$ in the bilateral ventral stream. In contrast, there are no consistent differences between aSTG and PTR, and between MTG and mSTS ($p > 0.05$). Towards the end of ventral trajectories, semantic representations become dominant in all subjects (bilateral PTR and MTG; $p < 0.05$; see Supp. Figs. A.1-5 and Supp. Fig A.6). These results indicate that representations in the ventral stream might be more

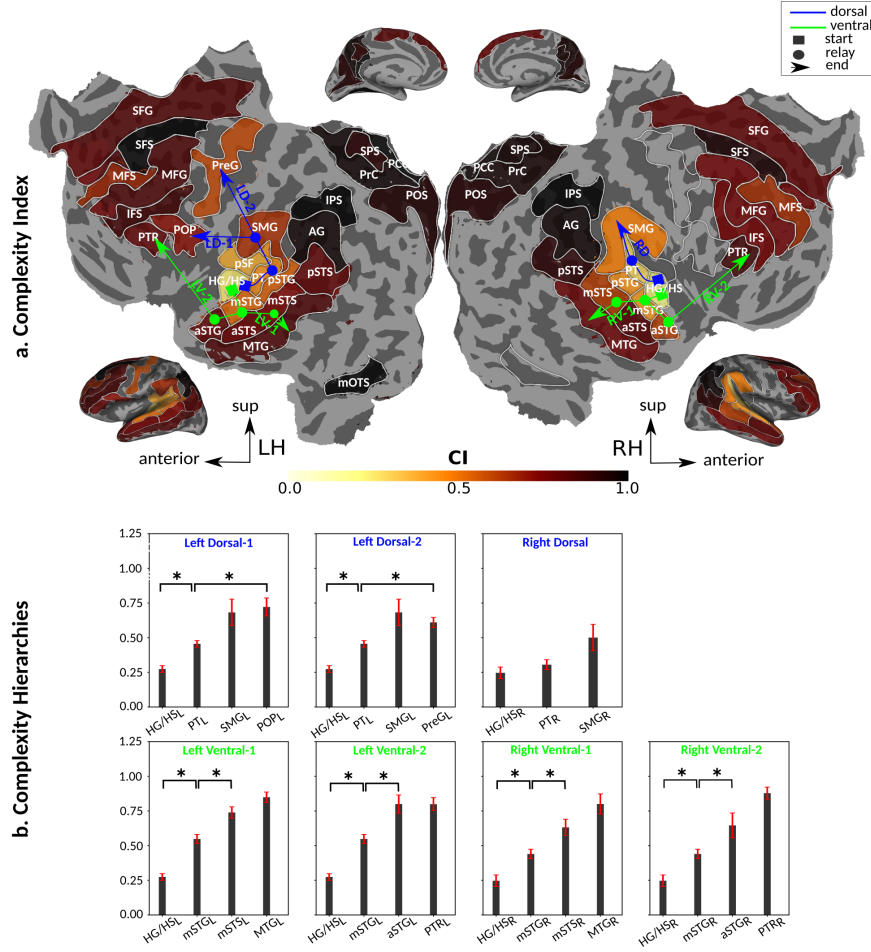


Figure 3.8: Cortical hierarchy of representational complexity. **a.** *Complexity index and auditory pathways.* To visualize how representational complexity distributes across cortex, cortical ROIs averaged across subjects are displayed on the flatmap of a representative subject (S4; see legend). To illustrate the gradients in across the hierarchy of speech processing, dorsal and ventral pathways are shown with blue and green lines, respectively: left dorsal-1 (LD-1), left dorsal-2 (LD-2) and right dorsal (RD), left ventral-1 (LV-1), left ventral-2 (LV-2), right ventral-1 (RV-1) and right ventral-2 (RV-2). Squares mark regions where pathways begin; arrows mark regions where pathways end; and circles mark relay regions in between. **b.** *Complexity hierarchies.* Bar plots display (mean $\pm$ sem across subjects) along LD-1, LD-2, RD, LV-1, LV-2, RV-1 and RV-2, shown in separate panels. Significant differences in CI between consecutive ROIs are marked with brackets ($p < 0.05$, bootstrap test). Significant gradients in CI are: $CI_{HG/HS} < CI_{PT} < CI_{POP}$ in LD-1, $CI_{HG/HS} < CI_{PT} < CI_{PreG}$ in LD-2, $CI_{HG/HS} < CI_{mSTG} < CI_{mSTS}$ in LV-1 and RV-1, and $CI_{HG/HS} < CI_{mSTG} < CI_{aSTG}$ in LV-2 and RV-2. These results suggest that hierarchical speech representations are systematically organized in multiple gradients across cortex.

symmetric across the two hemispheres compared to the dorsal stream.

3.3.1.2 Hemispheric asymmetries in speech representations

Prior studies report right lateralization in spectral representations [171, 172, 123, 173] and left lateralization in phonetic and semantic representations [157, 173]. Thus, it is likely that speech-related cortical regions are tuned for more complex speech features in the left versus the right hemisphere. To test this prediction, we compared CI between the left and right counterparts of each ROI, with consistent selectivity for speech features in both hemispheres in each individual subject (see Methods). Table 3.2 lists the results of this across-hemisphere comparison. No consistent hemispheric asymmetry in representational complexity is observed across cortex ($p > 0.05$) except in mSTG, with higher CI in left hemisphere ($p < 0.05$). These results suggest a slight left-hemispheric bias in representational complexity across intermediate stages during passive-listening tasks.

Table 3.2: Hemispheric asymmetries in CI and gAI.

	CI	gAI
HG/HS	N, 0.0	N, 0.1
PT	N, 0.1	N, 0.1
aSTG	N, 0.2	N, 0.1
mSTG	L, 0.1	L, 0.2
pSTG	N, 0.0	N, 0.2
aSTS	N, 0.0	N, 0.1
mSTS	N, 0.1	N, 0.2
pSTS	N, 0.1	N, 0.1
MTG	N, 0.1	N, 0.1
SMG	N, 0.2	N, 0.1
AG	N, 0.0	N, 0.0
IPS	N, 0.0	N, 0.2
SPS	N, 0.0	N, 0.0
PrC	N, 0.0	N, 0.0
PCC	N, 0.0	N, 0.1
POS	N, 0.0	N, 0.0
PTR	N, 0.1	N, 0.1
IFS	N, 0.1	N, 0.2
MFG	N, 0.1	N, 0.1
MFS	N, 0.0	N, 0.0
SFS	N, 0.0	N, 0.0
SFG	N, 0.0	N, 0.0

“R”, “L” and “N” indicate right-hemispheric, left-hemispheric and no-hemispheric dominance in indices respectively. The numbers show the effect-size of the significance tests, based on the difference between the left and right hemispheres. Only differences that are consistently significant in each individual subject are taken as significant ($p < 0.05$). ROIs are reported with at least 10 speech-selective voxels consistently in each hemisphere and in each subject ($p < 0.05$).

3.3.2 Attentional modulation of multi-level speech representations

Recent electrophysiology [123, 124, 126, 127, 129, 130] and neuroimaging studies [135, 136, 137] suggest that attention modulates cortical responses to speech during nonspatial or spatial cocktail-party tasks. However, less is known regarding the cortical distribution and strength of these modulations for the hierarchy of features involved in speech perception. Here, we systematically examined attentional modulations in low-level spectral, intermediate-level articulatory and high-level semantic representations during a diotic cocktail-party task. To do this, we recorded whole-brain BOLD responses while participants listened to temporally-overlaid spoken narratives from two different speakers and attended to either a male or female speaker in these two-speaker stories. We used the spectral, articulatory, and semantic models fit using passive-listening data to predict responses during the cocktail-party task. Since a voxel can represent information on both attended and unattended stimuli, response predictions were expressed as a convex combination of individual predictions for the attended and unattended story within each two-speaker story. Prediction scores were computed based on estimated responses as the combination weights were varied in $[0\ 1]$ (see Methods). Scores for the optimal combination model were compared against the scores from the individual models for attended and unattended stories. If the optimal combination model significantly outperforms the individual models, it indicates that the voxel represents information from both attended and unattended stimuli.

Figure 3.9 displays prediction scores of the spectral, articulatory, and semantic models as a function of the combination weight in representative ROIs, including HG, HS and PT. Scores based on only attended story (r_a), based on only the unattended story (r_u), and based on the optimal combination of the two (r_{max}) are marked. A diverse set of attentional effects are observed for each type of model. For the *spectral model* in left HG/HS, the optimal combination assigns matched weights to attended and unattended stories, and r_{max} is larger than r_a ($p < 10^{-4}$). This finding implies that spectral representations of the unattended story are mostly maintained; and there is no apparent bias towards the attended

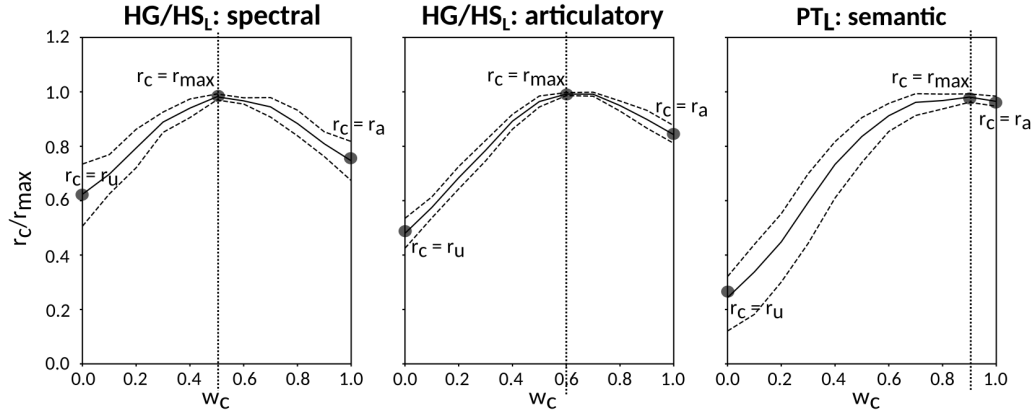


Figure 3.9: **Predicting cocktail-party responses.** Passive-listening models were tested during the cocktail-party task by predicting BOLD responses in the cocktail-party data. Since a voxel might represent information from both attended and unattended stimuli, response predictions were expressed as a convex combination of individual predictions for the attended and unattended story within each two-speaker story. Prediction scores were computed as the combination weights (w_c) were varied in [0 1] (see Methods). Prediction scores for a given model were averaged across speech-selective voxels within each ROI (r_c). The normalized scores of spectral, articulatory, and semantic models are displayed in several representative ROIs (HG/HS, HG/HS, and PT). Solid and dashed lines indicate mean and %95 confidence intervals across subjects. Scores based on only the attended story (r_a), based on only the unattended story (r_u), and based on the optimal combination of the two (r_{max}) are marked with circles. For the *spectral model* in left HG/HS, r_{max} is larger than r_a ($p < 10^{-4}$); and the optimal combination equally weighs attended and unattended stories. For the *articulatory model* in left HG/HS, r_a is larger than r_u ($p < 10^{-4}$), while r_{max} is greater than r_a ($p < 10^{-2}$). Besides, the optimal combination puts slightly higher weight to attended story than unattended story. For the *semantic model* in left PT, r_a is much higher than r_u ($p < 10^{-4}$), and the optimal combination puts much greater weight to attended story than unattended one. These representative results imply that attention may have divergent effects at various levels of speech representations across cortex.

story at spectral level in left HG/HS. For the *articulatory model* in left HG/HS, r_a is larger than r_u ($p < 10^{-4}$), while r_{max} is greater than r_a ($p < 10^{-2}$). Besides, the optimal combination gives slightly higher weight to the attended versus unattended story. This result suggests that attention moderately shifts articulatory representations in left HG/HS in favor of the attended stream such that articulatory representations of the unattended story are preserved to a degree. For the *semantic model* in left PT, r_a is much higher than r_u ($p < 10^{-4}$). Besides, the optimal combination assigns substantially higher weight to the attended story in this case. This finding indicates that attention strongly shifts semantic representations in left PT towards the attended stimulus. A simple inspection of these results suggests that attention may have distinct effects at various levels of speech representation across cortex. Hence, a detailed quantitative analysis is warranted to measure the effect of attention at each level.

3.3.2.1 Level-specific attentional modulations

To quantitatively assess the strength and direction of attentional modulations, we separately investigated the modulatory effects on spectral, articulatory, and semantic features across cortex. To measure modulatory effects at each feature level, a model-specific attention index (AI_m) was computed, reflecting the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1, 1]$; a positive index indicates selectivity modulation in favor of the attended stimulus, whereas a negative index indicates selectivity modulation in favor of the unattended stimulus. A value of zero indicates no modulation.

Figure 3.10a and Figure 3.11 display the attention index for spectral, articulatory, and semantic models across perisylvian and non-perisylvian ROIs, respectively (see Supp. Fig. A.7-11 for single-subject results). The modulation profiles of the ROIs are also visualized on the cortical flatmap, projecting articulatory, semantic, and spectral attention indices to the red, green, and blue channels of the RGB colormap as seen in Figure 3.10b (see Supp. Fig. A.12 for modulation profile flatmaps in individual subjects). Here we discuss the attention index for

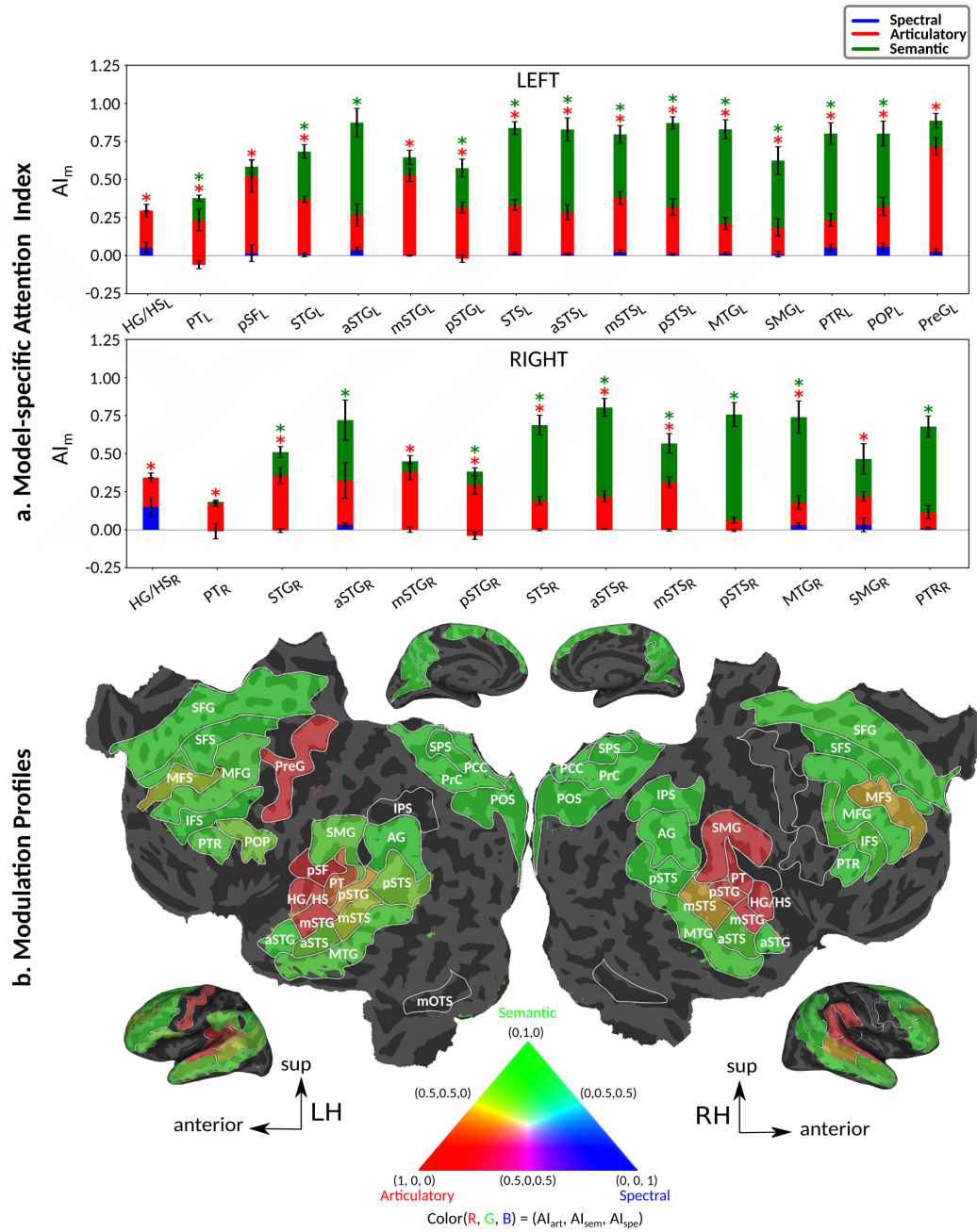


Figure 3.10: Caption next page

Figure 3.10: **Attentional modulation of multi-level speech representations.** **a.** *Model-specific attention indices.* A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI in perisylvian cortex, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across subjects), and their sum yields the overall modulation (see Figure 3.11 for non-perisylvian ROIs). Significantly positive indices are marked with * ($p < 0.05$, bootstrap test; see Supp. Figs. A.7-11 for attention indices of individual subjects). ROIs in the LH and RH are shown in top and bottom panels, respectively. These results show that selectivity modulations distribute broadly across cortex at the linguistic level (articulatory and semantic). **b.** *Attentional modulation profiles.* Modulation profiles averaged across subjects are displayed on the flattened cortical surface of a representative subject (S4). Significantly positive articulatory, semantic, and spectral attention indices are projected onto the red, green and blue channels of the colormap (see Methods). A progression in the level of speech representations dominantly modulated is apparent from HG/HS to MTG across bilateral temporal cortex (see Supp. Fig. A.12 for modulation profiles of individual subjects). Articulatory modulation is dominant in one end of the dorsal stream (left PreG), whereas semantic modulation becomes dominant in both ends of the ventral stream (bilateral PTR and MTG) ($p < 0.05$; see Supp. Figs. A.7-11 and A.12). On the other hand, semantic modulation is dominant in most of the higher-order regions in the parietal and frontal cortices consistently in all subjects (bilateral AG, SPS, PrC, PCC, POS, SFG, SFS, and PTR; left MFG; and right IPS) ($p < 0.05$; see Supp. Figs. A.7-11).

each model individually. *Spectral modulation* is not consistently significant in each subject across perisylvian ROIs ($p > 0.05$). On the other hand, moderate spectral modulation is found in right SFG consistently in all subjects ($p < 10^{-3}$). *Articulatory modulation* starts as early as HG/HS bilaterally ($p < 10^{-3}$). In the dorsal stream, it extends to PreG and POP in the left hemisphere (LH) and to SMG in the right hemisphere (RH; $p < 10^{-2}$); and it becomes dominant only in left PreG consistently in all subjects ($p < 0.05$). In the ventral stream, it extends to left PTR and bilateral MTG ($p < 10^{-2}$). Articulatory modulation is also found -albeit generally less strongly- in frontal regions (bilateral MFS; left MFG; and right IFS and SFG) consistently in all subjects ($p < 0.05$). In the dorsal stream, *semantic modulation* starts in PT and extends to POP in LH ($p < 10^{-2}$), whereas it is not apparent in the right dorsal stream ($p > 0.05$). In the ventral stream, semantic modulation starts in aSTG and mSTS bilaterally ($p < 0.05$). It extends to MTG and PTR, and becomes dominant in both ends of the bilateral

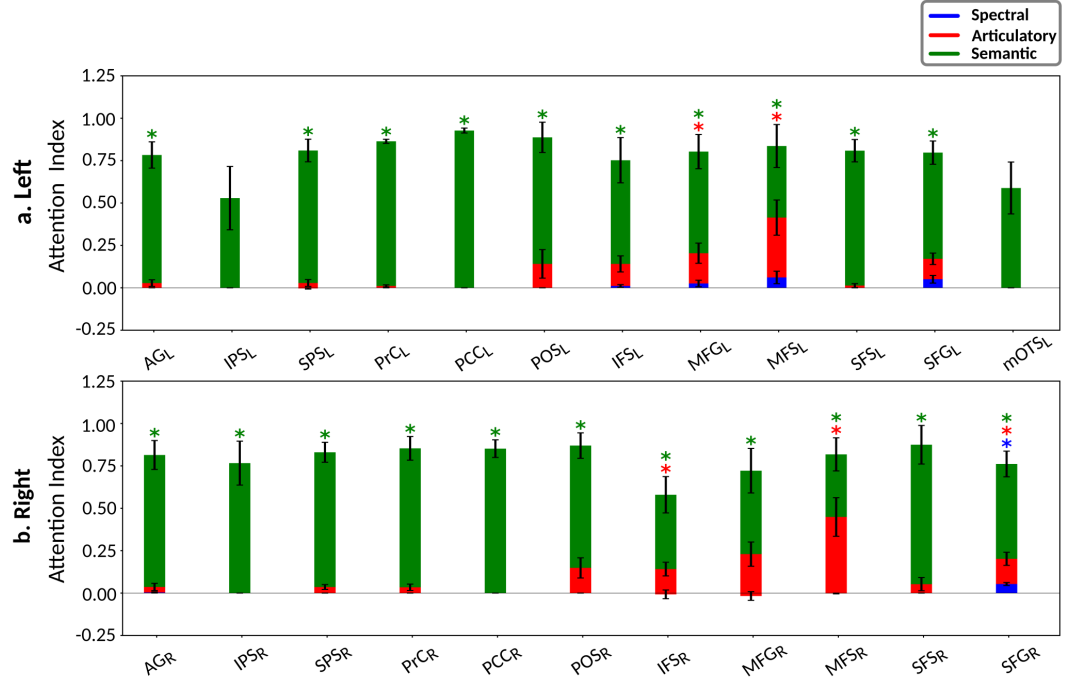


Figure 3.11: **Model-specific attention indices for non-perisylvian ROIs.** A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI in non-perisylvian cortex, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across subjects), and their sum yields the overall modulation. Significant indices are marked with * ($p < 0.05$, bootstrap test; see Supp. Figs. A.7-11 for attention indices of individual subjects). **a.** ROIs in *LH*. Modulations in left IPS and mOTS are not consistently significant in all subjects ($p > 0.05$). **b.** ROIs in *RH*. mOTS_R that did not have consistent speech selectivity in individual subjects was excluded (see Methods). These results indicate that selectivity modulations manifest primarily at the semantic level in most of the non-perisylvian ROIs.

ventral stream ($p < 0.05$). Lastly, semantic modulation is observed widespread across higher-order regions within frontal and parietal cortices consistently in all subjects ($p < 0.05$), with the exception of left IPS ($p < 0.05$). Taken together, these results suggest that attending to a target speaker alters articulatory and semantic representations broadly across cortex.

3.3.2.2 Global attentional modulations

It is commonly assumed that attentional effects grow stronger towards higher-order regions across the cortical hierarchy of speech [127, 130, 137]. Yet, a systematic examination of attentional modulation gradients across dorsal and ventral streams is lacking. To examine this issue, we measured overall attentional modulation in each region via a global attention index (gAI ; see Methods). Similar to the model-specific attention indices, a positive gAI indicates modulations in favor of the attended stimulus, and a negative gAI indicates modulations in favor of the unattended stimulus.

Dorsal stream. We first examined variation of gAI across the dorsal stream (left dorsal-1: $HG/HS_L \rightarrow PT_L \rightarrow (SMG_L) \rightarrow POP_L$, left dorsal-2: $HG/HS_L \rightarrow PT_L \rightarrow (SMG_L) \rightarrow PreG_L$, and right dorsal: $HG/HS_R \rightarrow PT_R \rightarrow SMG_R$) as shown in Figure 3.12. We find significant increase in gAI across the following left dorsal subtrajectories consistently in all subjects ($p < 0.05$; see Supp. Fig. A.14 for gradients in individual subjects): $gAI_{PT} < gAI_{SMG} < gAI_{POP}$ and $gAI_{PT} < gAI_{SMG} < gAI_{PreG}$. In contrast, we find no consistent gradient in the right dorsal stream ($p > 0.05$). These results suggest that attentional modulations grow progressively stronger across the dorsal stream in LH.

Ventral stream. We then examined variation of gAI across the ventral stream (left ventral-1: $HG/HS_L \rightarrow mSTG_L \rightarrow mSTS_L \rightarrow MTG_L$, left ventral-2: $HG/HS_L \rightarrow mSTG_L \rightarrow aSTG_L \rightarrow PTR_L$, right ventral-1: $HG/HS_R \rightarrow mSTG_R \rightarrow mSTS_R \rightarrow MTG_R$ and right ventral-2: $HG/HS_R \rightarrow mSTG_R \rightarrow aSTG_R \rightarrow PTR_R$), as shown in Figure 3.12. We find significant increase in gAI across the following subtrajectories consistently in all subjects ($p < 0.05$; see

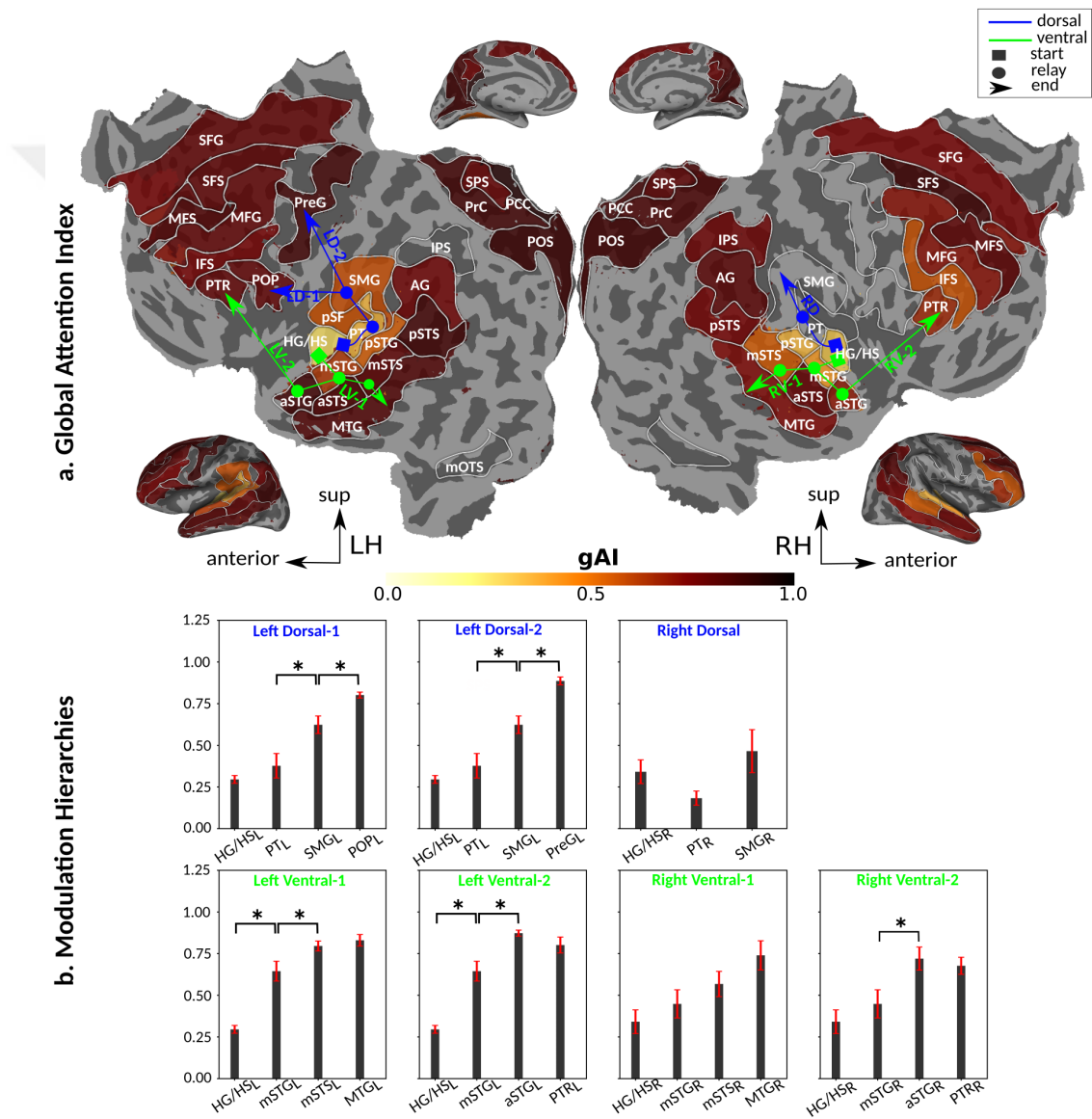


Figure 3.12: Caption next page

Figure 3.12: **Global attentional modulation.** **a.** *Global attention index.* To quantify overall modulatory effects on selectivity across all examined feature levels, global attentional modulation (gAI) was computed by summing spectral, articulatory, and semantic attention indices (see Methods). gAI is in the range of $[-1,1]$ and a value of zero indicates no modulation. Colors indicate significantly positive gAI averaged across subjects (see legend; see Supp. Fig. A.13 for bar plots of gAI across cortex). Dorsal and ventral pathways are shown with blue and green lines, respectively: left dorsal-1 (LD-1), left dorsal-2 (LD-2) and right dorsal (RD), left ventral-1 (LV-1), left ventral-2 (LV-2), right ventral-1 (RV-1) and right ventral-2 (RV-2). Squares mark regions where pathways begin; arrows mark regions where pathways end; and circles mark relay regions in gAI between. **b.** *Modulation hierarchies.* Bar plots display gAI (mean $\pm$ sem across subjects) along LD-1, LD-2, RD, LV-1, LV-2, RV-1 and RV-2, shown in separate panels. Significant differences in between consecutive ROIs are marked with brackets ($p < 0.05$, bootstrap test; see Supp. Fig. A.14 for single-subject results). Significant gradients in gAI are: $gAI_{PT} < gAI_{SMG} < gAI_{POP}$ in LD-1, $gAI_{PT} < gAI_{SMG} < gAI_{PreG}$ in LD-2, $gAI_{HG/HS} < gAI_{mSTG} < gAI_{mSTS}$ in LV-1, $gAI_{HG/HS} < gAI_{mSTG} < gAI_{aSTG}$ in LV-2, and $gAI_{mSTG} < gAI_{aSTG}$ in RV-2. In the left hemisphere, gAI gradually increases from early auditory regions to higher-order regions across the dorsal and ventral pathways. Similar patterns are also observed in the right hemisphere, although the gradients in gAI are less consistent across subjects.

Supp. Fig. A.14 for gradients in individual subjects): $gAI_{HG/HS} < gAI_{mSTG} < gAI_{aSTG}$ and $gAI_{HG/HS} < gAI_{mSTG} < gAI_{mSTS}$ in the left ventral stream, and $gAI_{mSTG} < gAI_{aSTG}$ in the right ventral stream. In contrast, we find no difference between aSTG and PTR bilaterally, between mSTS and MTG in the left ventral stream, and between HG/HS, mSTG, mSTS and MTG in the right ventral stream ($p > 0.05$). These results suggest that overall attentional modulations gradually increase across the ventral stream, and that the increases are more consistent in LH compared to RH.

Representational complexity versus attentional modulation. Visual inspection of Figure 3.8b and Figure 3.12b suggests that the subtrajectories with significant increases in CI and in gAI overlap largely in left ventral stream and partly in left dorsal stream. To quantitatively examine the overlap in left ventral stream, we analyzed the correlation between CI and gAI across the left-ventral subtrajectories where significant increases in CI are observed. We find significant correlations in $HG/HS \rightarrow mSTG \rightarrow mSTS$ and in $HG/HS \rightarrow mSTG \rightarrow aSTG$ ($r > 0.98$, bootstrap test, $p < 10^{-4}$) consistently in all subjects. In line with a recent study arguing for stronger attentional modulation and higher

representational complexity in STG compared to HG (O’Sullivan et al. 2019), our results indicate that attentional modulation increases towards higher-order regions as the representational complexity increases across the dorsal and ventral streams in LH (more apparent in ventral than dorsal stream).

Hemispheric asymmetries in attentional modulation. To assess potential hemispheric asymmetries in attentional modulation, we compared gAI between the left and right hemispheric counterparts of each ROI. This analysis was restricted to ROIs with consistent selectivity for speech features in both hemispheres in each individual subject (see Methods). Table 3.2 lists the results of the across-hemisphere comparison. No consistent hemispheric asymmetry is found across cortex with the exception of mSTG having a left-hemispheric bias consistently in all subjects ($p < 0.05$). These results indicate that there is mild lateralization in attentional modulation of intermediate-level speech features.

3.3.3 Cortical representation of unattended speech

An important question regarding multi-speaker speech perception is to what extent unattended stimuli are represented in cortex. To address this question, here we investigated spectral, articulatory, and semantic representations of unattended stories during the cocktail-party task. We reasoned that if significant information about unattended speech is represented in a brain region, then features of unattended speech should explain significant variance in measured BOLD responses. To test this, we compared the prediction score of a combination model comprising the features of both attended and unattended stories (optimal convex combination) against the prediction score of an individual model comprising only the features of the attended story (see Methods). If the combination model significantly outperforms the individual model in an ROI, then the corresponding features of unattended speech are significantly represented in that ROI.

Figure 3.13 displays model performance when responses are predicted based on speech features from the attended story alone, and when they are instead

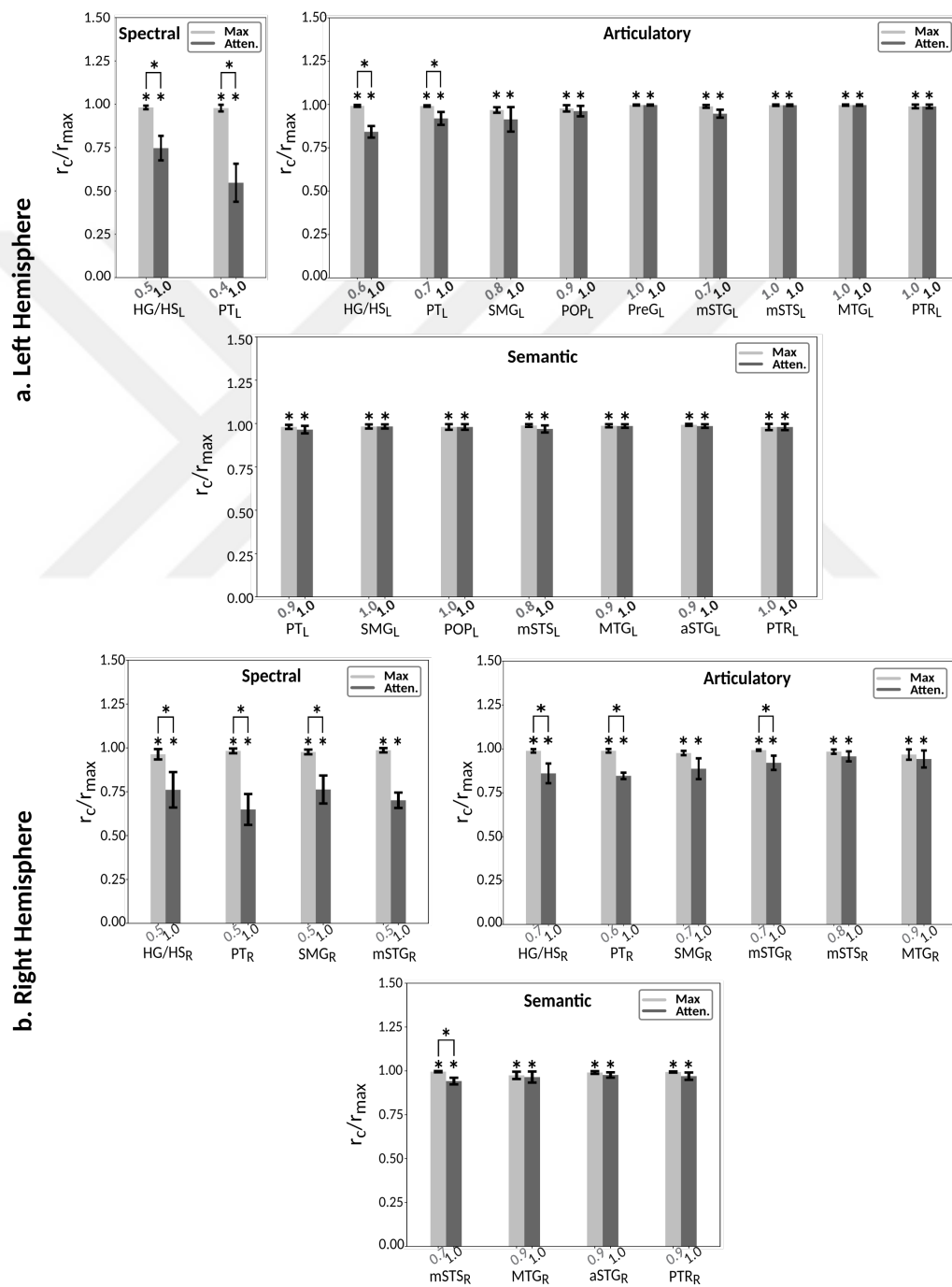


Figure 3.13: Caption next page

Figure 3.13: Representation of unattended speech. Passive-listening models were tested on cocktail-party data to assess representation of unattended speech during the cocktail-party task. Prediction scores were calculated separately for a combination model comprising features of both attended and unattended stories (r_{max} : optimal convex combination) and an individual model only comprising features of the attended story (r_a). Significant difference in prediction between the two models is an indication that BOLD responses carry significant information on unattended speech. Bar plots display normalized prediction scores (mean $\pm$ sem across subjects; combination model in light gray and individual model in gray). Significant scores are marked with * ($p < 10^{-4}$, bootstrap test; see Supp. Figs. A.15-19 for single-subject results), and significant differences are marked with brackets ($p < 0.05$). Prediction scores are displayed for ROIs in the dorsal and ventral streams, with significant selectivity for given model features. **a. Left hemisphere.** Spectral representations of unattended speech extend up to PT across the dorsal stream ($HG/HS \rightarrow PT \rightarrow SMG \rightarrow (POP, PreG)$) and are constrained to HG/HS across the ventral stream ($HG/HS \rightarrow mSTG \rightarrow mSTS \rightarrow MTG$) and ($HG/HS \rightarrow mSTG \rightarrow aSTG \rightarrow PTR$). Articulatory representations of unattended speech extend up to PT across the dorsal stream and are constrained to HG/HS across the ventral stream. No significant semantic representation is apparent. **b. Right hemisphere.** Spectral representations of unattended speech extend up to SMG across the dorsal stream and are constrained to HG/HS across the ventral stream. Articulatory representations of unattended speech extend up to PT across the dorsal stream, and up to mSTG across the ventral stream. Semantic representations are found only in mSTS. These results suggest that processing of unattended speech is not constrained at spectral level but extends to articulatory and semantic level.

predicted based on the optimally combined features from the attended and unattended stories. Results are shown for each ROI along the dorsal and ventral streams and in the left and right hemispheres (see Supp. Figs. A.15-19 for single-subject results). Along the left ($HG/HS \rightarrow PT \rightarrow SMG \rightarrow (POP, PreG)$) and right ($HG/HS \rightarrow PT \rightarrow SMG$) dorsal stream, spectral features of unattended speech are represented up to PT in LH and up to SMG in RH ($p < 0.01$), articulatory features are represented bilaterally up to PT ($p < 0.05$), whereas no semantic representation is apparent ($p > 0.05$). Along the left ventral stream ($HG/HS \rightarrow mSTG \rightarrow mSTS \rightarrow MTG$) and ($HG/HS \rightarrow mSTG \rightarrow aSTG \rightarrow PTR$), spectral and articulatory features are represented in HG/HS ($p < 10^{-4}$), again with no semantic representation ($p > 0.05$). In the right ventral stream ($HG/HS \rightarrow mSTG \rightarrow mSTS \rightarrow MTG$) and ($HG/HS \rightarrow mSTG \rightarrow aSTG \rightarrow PTR$), spectral features are represented in HG/HS; articulatory features are represented up to mSTG ($p < 0.05$); and semantic features are represented only in mSTS ($p < 0.05$). These results indicate

that cortical representations of unattended speech in multi-speaker environments extend from the spectral to the semantic level, albeit semantic representations are constrained to right parabelt auditory cortex (mSTS). Furthermore, representations of unattended speech are more broadly spread across the right hemisphere. Note that prior studies have reported response correlations and anatomical overlap between these belt/parabelt auditory regions and the reorienting attention system in the right-hemisphere [174, 175, 176]. Therefore, relatively broader representations of unattended speech in the right hemisphere might facilitate distractor detection and filtering during auditory attention tasks.

3.4 Discussion

In this study, we investigated the effects of auditory attention on multi-level speech representations across cortex during a diotic cocktail-party task with naturalistic stimuli composed of spoken narratives. To assess baseline selectivity for multi-level speech features, we first fit spectral, articulatory, and semantic models using responses recorded during passive listening. We then quantified the complexity of intrinsic representations in each brain region. Next, we used fit models that reflect baseline selectivity for speech features to assess attentional modulation of speech representations. To do this, responses predicted using stimulus features of attended and unattended stories were compared with responses recorded during the cocktail-party task. This study is among the first to quantitatively characterize attentional modulations in multi-level speech representations of attended and unattended stimuli across speech-related cortex.

3.4.1 Speech representations

Prior studies using controlled auditory stimuli have suggested a hierarchical organization of speech representations across cortex [177, 17, 178, 179, 157]. Recent studies using natural speech stimuli corroborate this view where downstream regions are notably selective for more complex speech features than early regions

in speech-related cortex [119, 120, 129]. Motivated by these results, here we quantitatively examined gradients in complexity of speech representations across speech-related cortex. Overall, we find that representational complexity increases gradually across the left dorsal and bilateral ventral streams.

An important question in neurolinguistics is the precise functional roles of the dorsal and ventral streams for natural speech representations in the human brain. During passive listening of natural stories, we find that dominant articulatory selectivity manifests (PreG) in one end of the dorsal stream whereas semantic selectivity manifests in both ends (MTG and PTR) of the ventral stream. These results confirm the view that the dorsal stream is implicated in sound-to-articulation mapping, whereas the ventral stream is primarily implicated in sound-to-meaning mapping [18, 19, 20, 162] .

Linguistic representations are commonly considered to be left lateralized while acoustic representations (spectral) are suggested to be right lateralized [171, 123, 157, 173]. Here we quantitatively assessed hemispheric asymmetries in multi-level speech representations along the dorsal and ventral streams. We report a left-hemispheric dominance in the dorsal stream that becomes relatively more apparent towards the end stages (PreG and POP). However, regions along the ventral stream have similar selectivity profiles in both hemispheres, with a slight left-hemispheric bias in their representational complexity in intermediate stages. Hence, our results indicate a left-lateralized organization in the dorsal stream and a predominantly bilateral organization in the ventral stream [18].

3.4.2 Attentional modulations

The effects of auditory attention on cortical responses have been primarily examined in the literature using controlled stimuli such as simple tones, melodies and isolated syllables or words [93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 180, 181, 182]. As such, less is known regarding how attention alters hierarchical representations of natural speech. Recent studies on this topic have mainly reported attentional modulations of low-level speech representations comprising

speech-envelope and spectrogram features in early auditory and higher-order regions during the cocktail-party task [123, 124, 125, 127, 128, 131]. Going beyond, here we have explored attentional modulations spanning from low-level spectral to high-level semantic features. While our results indicate that attentional modulations for articulatory and semantic representations distribute broadly across cortex, we find no consistent modulations for spectral representations in speech-related regions. Note that speech envelope and spectrogram features in natural speech carry intrinsic information about linguistic features including syllabic boundaries and articulatory features [183, 118]. These stimulus correlations can render it challenging to dissociate unique selectivity for articulatory versus spectral features. To minimize biases from potential stimulus correlations, here we leveraged a decorrelation procedure to obtain orthogonal spectral, articulatory, and semantic feature matrices for the stimulus. Therefore, the distinct modeling procedures for natural speech features might have contributed to the disparities between the current and previous studies on the existence of spectral modulations.

An important question regarding auditory attention is how the strength of attentional effects are distributed across cortex. A common view is that attentional modulations grow relatively stronger towards later stages of processing [127]. Recent studies support this view by reporting bilaterally stronger modulations in frontal versus temporal cortex [137] and in non-primary versus primary auditory cortex [130]. Adding to this body of evidence, we further show that attentional modulations gradually increase across the dorsal and ventral streams in the left hemisphere, as the complexity of speech representations grow. While a similar trend is observed across the right hemisphere, gradients in attentional modulation are less consistent in right belt and parabelt auditory regions including PT. Furthermore, attentional modulations are weaker in the right versus left hemisphere within these regions. Note that belt and parabelt regions are suggested to be connected to the right temporo-parietal junction (TPJ) during selective listening [176]. TPJ is one of the central nodes in the reorienting attention system that monitors salient events to filter out distractors and help maintaining focused attention [184, 174, 175]. Hence less consistent gradients and relatively weaker attentional modulations in right belt and parabelt auditory regions might

suggest a functional role for these regions in detecting salient events within the unattended stream during selective listening tasks.

Another central question regarding mechanisms of selective attention in a multi-speaker environment is how attentional modulations distribute across well-known dorsal and ventral streams. The dorsal stream that hosts articulatory representations is commonly considered to be involved in sound-to-articulation mapping [18, 19, 20, 162]. The motor-theory of speech perception suggests that dorsal articulatory representations carry information about articulatory gestures of the speaker to facilitate the listener’s comprehension [185, 17, 186, 170, 187]. Recent studies support this account by reporting enhanced activity in precentral gyrus and premotor cortex during challenging listening conditions [188, 136, 189]. In accordance with the motor theory of speech perception, here we find predominant articulatory selectivity and modulation due to selective listening in one end of the dorsal stream (PreG). These articulatory modulations might serve to increase sensitivity to the target speaker’s gestures to facilitate speech comprehension during difficult cocktail-party tasks [136].

In contrast to the dorsal stream, the ventral stream has been implicated in sound-to-meaning mapping [18, 19, 20, 162]. Compatible with this functional role, the ventral stream is suggested to transform acoustic representations of linguistic stimuli into object-based representations [158]. Here, we find that representational complexity of the speech features gradually increases across bilateral ventral stream, and semantic representations become dominant at the ends of it (bilateral PTR and MTG). In addition, speech level of attentional modulation also progresses across the ventral stream, and strong and predominant semantic modulations manifest towards later stages. Hence, the ventral stream might serve as a stage for interplay between bottom-up processing and top-down attentional modulation to gradually form auditory objects during selective listening [158, 22, 190].

3.4.3 Representation of the unattended speech

Whether unattended speech is represented in cortex during selective listening and if so, at what feature levels its representations are maintained are crucial aspects of auditory attention. Behavioral accounts suggest that unattended speech is primarily represented at the acoustic level [25, 91]. Corroborating these accounts, recent electrophysiology studies have identified acoustic representations of unattended speech localized to auditory cortex [123, 124, 127, 128, 129, 130, 131]. In contrast, here we find that acoustic representations of unattended speech extend beyond the auditory cortex as far as SMG in the right dorsal stream. Because SMG partly overlaps with the reorienting attention system, unattended speech representations in this region might contribute to filtering of distractors during the cocktail-party task [174, 175].

A more controversial issue is whether unattended speech representations carry information at the linguistic level [90, 191, 192, 109, 106]. Prior studies on this issue are split between those suggesting the presence [136, 142] versus absence [193, 129] of linguistic representations. Here, we find that articulatory representations of unattended speech extend up to belt/parabelt auditory areas in the bilateral dorsal stream and the right ventral stream. We further find semantic representation of unattended speech in the right ventral stream (mSTS). These linguistic representations of unattended speech are naturally weaker than those of attended speech, and they are localized to early-to-intermediate stages of auditory processing. Our findings suggest that unattended speech is represented at the linguistic level prior to entering the broad semantic system where full selection of the attended stream occurs [194, 195, 196, 197, 198, 109, 199]. Overall, these linguistic representations might serve to direct exogenous triggering of attention to salient features in unattended speech [86, 87, 200, 201, 90, 109]. Meanwhile, attenuated semantic representations in the ventral stream might facilitate semantic priming of the attended stream by relevant information in the unattended stream [202, 90, 203].

Chapter 4

Conclusion

The main goal of this thesis was to find multilevel speech representations and to explore how attention affects these representations during a naturalistic cocktail-party task. Our results from passive-listening experiment indicate that linguistic representations of the speech are broadly distributed across cortex, while acoustic representations are mainly constrained to the early auditory cortex. Besides, the complexity of the representations gradually increases across both dorsal and ventral streams, suggesting a hierarchical organization of speech in the brain. Our results also confirm the view that dorsal stream is implicated in sound-to-articulation mapping, while ventral stream is mainly implicated in sound-to-meaning mapping.

Our results from cocktail-party experiment indicate that attention during a diotic cocktail-party task gradually selects attended over unattended speech across both dorsal and ventral processing pathways. This selection is mediated by representational modulations for linguistic features but not for acoustic features. Despite broad attentional modulations in favor of the attended stream, we still find that unattended speech is represented up to linguistic level in the regions that overlap with the reorienting attention system. These linguistic representations of unattended speech might facilitate attentional reorienting and filtering during natural speech perception. Overall, our findings provide comprehensive

insights on attentional mechanisms that underlie the ability to selectively listen to a desired speaker in noisy multi-speaker environments.

4.1 Future Directions

Reliability of statistical assessments in neuroimaging depends on two main factors: sample size and amount of data collected per subject. Given experimental constraints, it is difficult to increase both factors in a single study. In this unavoidable trade-off, a common practice in fMRI studies is to collect a relatively limited dataset from more subjects. This practice prioritizes across-subject variability over within-subject variability, at the expense of individual-subject results. Diverting away from this practice, we collected a larger amount of data per subject to give greater focus to reliability in single subjects. This choice is motivated by the central aims of the voxel-wise modeling (VM) approach. The VM framework aims to sensitively measure tuning profiles of single voxels in individual subjects. For the natural speech perception experiments conducted here, the tuning profiles were characterized over three separate high-dimensional spaces containing hundreds of acoustic and linguistic features. To maximize sensitivity of VM models, we conducted extensive experiments in each individual subject to increase the amount and diversity of fMRI data collected. This design enhanced the quality of resulting VM models and reliability of individual-subject results. Indeed, here we find highly uniform results across individual subjects, suggesting that the reported effects are highly robust. That said, across subject and across language variability might occur in diverse, multi-lingual cohorts. Assessment of attentional effects on speech representations in these broader populations remains important future work.

In the current study, subjects were presented continuous natural speech stimuli. In the passive-listening task, they were instructed to vigilantly listen to the presented story. Our analyses reveal that BOLD responses in large swaths of

language-related areas can be significantly predicted by voxel-wise models comprising spectral, articulatory and semantic speech features. Moreover, the spectral, articulatory and semantic representations mapped in single subjects are highly consistent across subjects. Therefore, these results suggest that the participants performed reasonably well in active listening of single-speaker stories. In the cocktail-party experiment, subjects were instead instructed to attentively listen to one of two speakers. Our analyses in this case reveal broad attentional modulations in representation of semantic information across cortex, in favor of the target speaker. Semantic features of natural speech show gradual variation across time compared to low-level spectral information. Therefore, this finding suggests that subjects also performed well during the sustained attention tasks in the cocktail-party task. That said, we cannot rule out momentary shifts in attention away from the target speaker. If momentary shifts towards the unattended speaker are frequent, they might increase the proportion of unattended speech information that BOLD responses carry. In turn, this might have contributed to the strength of unattended speech representations that we measured during the cocktail-party task. Post-scan questionnaires that assess participants' comprehension of attended and unattended stories are a common control for task execution [137]. However, post-scan memory controls cannot guarantee lack of momentary attention shifts that typically last less than 200 ms [204]. On the other hand, implementing frequent controls during the scan itself would disrupt the naturalistic experiment flow and efficiency. It is therefore challenging to experimentally monitor momentary attentional shifts [109]. To assess the influence of momentary shifts during sustained-attention tasks, future studies are warranted leveraging more controlled speech stimuli with systematic variations in the salience and task relevance of non-target stimuli [174, 205, 109].

In the current study, we find that attention strongly alters semantic representations in favor of the target stream across frontal and parietal cortices. This is in alignment with previous fMRI studies that found attentional response modulations in frontal and parietal regions [134, 135, 137]. That said, an important question is whether these modulations predominantly reflect enhanced bottom-up processing of attended speech or top-down attentional control signals [184, 181].

Note that we find broad semantic representations across frontal and parietal cortices during the passive-listening experiment, in the absence of any demanding attentional tasks. Furthermore, a recent study suggests that various semantic categories are differentially represented in these higher-level cortical regions [143]. Taken together, these findings imply that semantic modulations in fronto-parietal regions can be partly attributed to bottom-up effects. Yet, it is challenging to disentangle bottom-up and top-down contributions in fMRI studies due to the inherently limited temporal resolution. Future studies are warranted to shed light on this issue by combining the spatial sampling capability of fMRI with high temporal resolution of electrophysiology methods [79, 119].

Bibliography

- [1] F. A. Azevedo, L. R. Carvalho, L. T. Grinberg, J. M. Farfel, R. E. Ferretti, R. E. Leite, W. J. Filho, R. Lent, and S. Herculano-Houzel, “Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain,” *Journal of Comparative Neurology*, vol. 513, no. 5, pp. 532–541, 2009.
- [2] S. Herculano-Houzel, “The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost,” *Proceedings of the National Academy of Sciences*, vol. 109, no. Supplement 1, pp. 10661–10668, 2012.
- [3] R. C. Berwick, A. D. Friederici, N. Chomsky, and J. J. Bolhuis, “Evolution, brain, and the nature of language,” *Trends in cognitive sciences*, vol. 17, no. 2, pp. 89–98, 2013.
- [4] A. L. Benton, “Contributions to aphasia before broca,” *Cortex*, vol. 1, no. 3, pp. 314–327, 1964.
- [5] P. Tremblay and A. S. Dick, “Broca and wernicke are dead, or moving past the classic model of language neurobiology,” *Brain and language*, vol. 162, pp. 60–71, 2016.
- [6] N. Geschwind, “The organization of language and the brain,” *Science*, vol. 170, no. 3961, pp. 940–944, 1970.
- [7] J. Crinion, A. Holland, D. Copland, C. K. Thompson, and A. E. Hillis, “Quantifying brain lesions in neuroimaging research examining language recovery after stroke,” *NeuroImage*, vol. 73, p. 208, 2013.

- [8] D. Mirman, Q. Chen, Y. Zhang, Z. Wang, O. K. Faseyitan, H. B. Coslett, and M. F. Schwartz, “Neural organization of spoken language revealed by lesion–symptom mapping,” *Nature communications*, vol. 6, no. 1, pp. 1–9, 2015.
- [9] J. R. Binder, R. H. Desai, W. W. Graves, and L. L. Conant, “Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies,” *Cerebral cortex*, vol. 19, no. 12, pp. 2767–2796, 2009.
- [10] C. J. Price, “The anatomy of language: a review of 100 fmri studies published in 2009,” *Annals of the new York Academy of Sciences*, vol. 1191, no. 1, pp. 62–88, 2010.
- [11] C. J. Price, “A review and synthesis of the first 20 years of pet and fmri studies of heard speech, spoken language and reading,” *Neuroimage*, vol. 62, no. 2, pp. 816–847, 2012.
- [12] M. Catani, D. K. Jones, and D. H. Ffytche, “Perisylvian language networks of the human brain,” *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 57, no. 1, pp. 8–16, 2005.
- [13] M. Catani and M. T. De Schotten, “A diffusion tensor imaging tractography atlas for virtual in vivo dissections,” *cortex*, vol. 44, no. 8, pp. 1105–1132, 2008.
- [14] S. M. Gierhan, “Connections for auditory language in the human brain,” *Brain and language*, vol. 127, no. 2, pp. 205–221, 2013.
- [15] A. D. Friederici and S. M. Gierhan, “The language network,” *Current opinion in neurobiology*, vol. 23, no. 2, pp. 250–254, 2013.
- [16] A. D. Friederici, “White-matter pathways for speech and language processing,” *Handbook of clinical neurology*, vol. 129, pp. 177–186, 2015.
- [17] G. Hickok and D. Poeppel, “Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language,” *Cognition*, vol. 92, no. 1-2, pp. 67–99, 2004.

- [18] G. Hickok and D. Poeppel, “The cortical organization of speech processing,” *Nature reviews neuroscience*, vol. 8, no. 5, pp. 393–402, 2007.
- [19] J. P. Rauschecker and S. K. Scott, “Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing,” *Nature neuroscience*, vol. 12, no. 6, pp. 718–724, 2009.
- [20] A. D. Friederici, “The brain basis of language processing: from structure to function,” *Physiological reviews*, vol. 91, no. 4, pp. 1357–1392, 2011.
- [21] B. G. Shinn-Cunningham and V. Best, “Selective attention in normal and impaired hearing,” *Trends in amplification*, vol. 12, no. 4, pp. 283–299, 2008.
- [22] B. Shinn-Cunningham, V. Best, and A. K. Lee, “Auditory object formation and selection,” in *The auditory system at the cocktail party*, pp. 7–40, Springer, 2017.
- [23] J. H. McDermott, “The cocktail party problem,” *Current Biology*, vol. 19, no. 22, pp. R1024–R1027, 2009.
- [24] C. W. Di Fu, G. Yang, M. Kerzel, W. Nan, P. Barros, H. Wu, X. Liu, and S. Wermter, “What can computational models learn from human selective attention? a review from an audiovisual unimodal and crossmodal perspective,” *Frontiers in integrative neuroscience*, vol. 14, 2020.
- [25] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.
- [26] K. Alho, T. Rinne, T. J. Herron, and D. L. Woods, “Stimulus-dependent activations and attention-related modulations in the auditory cortex: a meta-analysis of fmri studies,” *Hearing research*, vol. 307, pp. 29–41, 2014.
- [27] L. M. Miller, “Neural mechanisms of attention to speech,” in *Neurobiology of Language*, pp. 503–514, Elsevier, 2016.

- [28] M. Shahdloo, *Optimization and machine learning in MRI: applications in rapid MR image reconstruction and encoding models of cortical representations*. PhD thesis, Bilkent University, 2020.
- [29] L. S. Hamilton and A. G. Huth, “The revolution will not be controlled: natural stimuli in speech neuroscience,” *Language, Cognition and Neuroscience*, vol. 35, no. 5, pp. 573–582, 2020.
- [30] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant, “Encoding and decoding in fmri,” *Neuroimage*, vol. 56, no. 2, pp. 400–410, 2011.
- [31] K. Alho and V. A. Vorobyev, “Brain activity during selective listening to natural speech,” *Front Biosci*, vol. 12, pp. 3167–3176, 2007.
- [32] M. A. Lindquist *et al.*, “The statistical analysis of fmri data,” *Statistical science*, vol. 23, no. 4, pp. 439–464, 2008.
- [33] M. A. Lindquist and T. D. Wager, “Principles of functional magnetic resonance imaging,” *Handbook of neuroimaging data analysis*, pp. 3–48, 2014.
- [34] N. K. Logothetis, “What we can do and what we cannot do with fmri,” *Nature*, vol. 453, no. 7197, pp. 869–878, 2008.
- [35] T. Çukur, N. K. Bangerter, and D. G. Nishimura, “Enhanced spectral shaping in steady-state free precession imaging,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1216–1223, 2007.
- [36] T. Çukur and D. G. Nishimura, “Fat-water separation with alternating repetition time balanced ssfp,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 60, no. 2, pp. 479–484, 2008.
- [37] T. Çukur and D. G. Nishimura, “Multiple repetition time balanced steady-state free precession imaging,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 62, no. 1, pp. 193–204, 2009.

- [38] T. Çukur, J. H. Lee, N. K. Bangerter, B. A. Hargreaves, and D. G. Nishimura, “Non-contrast-enhanced flow-independent peripheral mr angiography with balanced ssfp,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 61, no. 6, pp. 1533–1539, 2009.
- [39] T. Çukur, M. Yamada, W. R. Overall, P. Yang, and D. G. Nishimura, “Positive contrast with alternating repetition time ssfp (parts): A fast imaging technique for spio-labeled cells,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 63, no. 2, pp. 427–437, 2010.
- [40] K. Keskin, U. Yılmaz, and T. Çukur, “Constrained ellipse fitting for efficient parameter mapping with phase-cycled bssfp mri,” *arXiv preprint arXiv:2106.03239*, 2021.
- [41] D. J. Larkman, J. V. Hajnal, A. H. Herlihy, G. A. Coutts, I. R. Young, and G. Ehnholm, “Use of multicoil arrays for separation of signal from multiple slices simultaneously excited,” *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 13, no. 2, pp. 313–317, 2001.
- [42] S. Moeller, E. Yacoub, C. A. Olman, E. Auerbach, J. Strupp, N. Harel, and K. Ugurbil, “Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri,” *Magnetic resonance in medicine*, vol. 63, no. 5, pp. 1144–1153, 2010.
- [43] E. Bıyık, K. Keskin, S. UH Dar, A. Koç, and T. Çukur, “Factorized sensitivity estimation for artifact suppression in phase-cycled bssfp mri,” *NMR in Biomedicine*, vol. 33, no. 4, p. e4228, 2020.
- [44] E. Biyik, E. Ilicak, and T. Çukur, “Reconstruction by calibration over tensors for multi-coil multi-acquisition balanced ssfp imaging,” *Magnetic resonance in medicine*, vol. 79, no. 5, pp. 2542–2554, 2018.

- [45] B. Quist, B. A. Hargreaves, T. Cukur, G. R. Morrell, G. E. Gold, and N. K. Bangerter, “Simultaneous fat suppression and band reduction with large-angle multiple-acquisition balanced steady-state free precession,” *Magnetic resonance in medicine*, vol. 67, no. 4, pp. 1004–1012, 2012.
- [46] T. Cukur, J. M. Santos, J. M. Pauly, and D. G. Nishimura, “Variable-density parallel imaging with partially localized coil sensitivities,” *IEEE transactions on medical imaging*, vol. 29, no. 5, pp. 1173–1181, 2010.
- [47] T. Çukur, M. Lustig, and D. G. Nishimura, “Multiple-profile homogeneous image combination: application to phase-cycled ssfp and multicoil imaging,” *Magnetic resonance in medicine*, vol. 60, no. 3, pp. 732–738, 2008.
- [48] E. Ilicak, L. K. Senel, E. Biyik, and T. Çukur, “Profile-encoding reconstruction for multiple-acquisition balanced steady-state free precession imaging,” *Magnetic resonance in medicine*, vol. 78, no. 4, pp. 1316–1329, 2017.
- [49] A. Koç, B. Bartan, E. Gundogdu, T. Çukur, and H. M. Ozaktas, “Sparse representation of two-and three-dimensional images with fractional fourier, hartley, linear canonical, and haar wavelet transforms,” *Expert Systems with Applications*, vol. 77, pp. 247–255, 2017.
- [50] E. Ilicak, S. Cetin, E. Bulut, K. K. Oguz, E. U. Saritas, G. Unal, and T. Çukur, “Targeted vessel reconstruction in non-contrast-enhanced steady-state free precession angiography,” *NMR in Biomedicine*, vol. 29, no. 5, pp. 532–544, 2016.
- [51] T. Çukur, “Spectrally selective imaging with wideband balanced steady-state free precession mri,” *Magnetic resonance in medicine*, vol. 75, no. 3, pp. 1132–1141, 2016.
- [52] T. Çukur, J. M. Santos, D. G. Nishimura, and J. M. Pauly, “Varying kernel-extent gridding reconstruction for undersampled variable-density spirals,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 59, no. 1, pp. 196–201, 2008.

- [53] E. Kopanoglu, A. Güngör, T. Kilic, E. U. Saritas, K. K. Oguz, T. Çukur, and H. E. Güven, “Simultaneous use of individual and joint regularization terms in compressive sensing: Joint reconstruction of multi-channel multi-contrast mri acquisitions,” *NMR in Biomedicine*, vol. 33, no. 4, p. e4247, 2020.
- [54] S. Ilbey, C. B. Top, A. Güngör, T. Çukur, E. U. Saritas, and H. E. Güven, “Fast system calibration with coded calibration scenes for magnetic particle imaging,” *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2070–2080, 2019.
- [55] L. K. Senel, T. Kilic, A. Gungor, E. Kopanoglu, H. E. Guven, E. U. Saritas, A. Koc, and T. Cukur, “Statistically segregated k-space sampling for accelerating multiple-acquisition mri,” *IEEE transactions on medical imaging*, vol. 38, no. 7, pp. 1701–1714, 2019.
- [56] M. Shahdloo, E. Ilıcak, M. Tofghi, E. U. Saritas, A. E. Çetin, and T. Çukur, “Projection onto epigraph sets for rapid self-tuning compressed sensing mri,” *IEEE transactions on medical imaging*, vol. 38, no. 7, pp. 1677–1689, 2018.
- [57] B. Gözcü, R. K. Mahabadi, Y.-H. Li, E. Ilıcak, T. Cukur, J. Scarlett, and V. Cevher, “Learning-based compressive mri,” *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1394–1406, 2018.
- [58] T. Çukur, “Accelerated phase-cycled ssfp imaging with compressed sensing,” *IEEE transactions on medical imaging*, vol. 34, no. 1, pp. 107–115, 2014.
- [59] T. Çukur, A. Shimakawa, H. Yu, B. A. Hargreaves, B. S. Hu, D. G. Nishimura, and J. H. Brittain, “Magnetization-prepared ideal bssfp: a flow-independent technique for noncontrast-enhanced peripheral angiography,” *Journal of Magnetic Resonance Imaging*, vol. 33, no. 4, pp. 931–939, 2011.
- [60] T. Çukur, M. Lustig, E. U. Saritas, and D. G. Nishimura, “Signal compensation and compressed sensing for magnetization-prepared mr angiography,” *IEEE transactions on medical imaging*, vol. 30, no. 5, pp. 1017–1027, 2011.

- [61] T. Cukur, M. Lustig, and D. G. Nishimura, “Improving non-contrast-enhanced steady-state free precession angiography with compressed sensing,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 61, no. 5, pp. 1122–1131, 2009.
- [62] M. Yurt, S. U. Dar, A. Erdem, E. Erdem, K. K. Oguz, and T. Çukur, “must-gan: Multi-stream generative adversarial networks for mr image synthesis,” *Medical Image Analysis*, vol. 70, p. 101944, 2021.
- [63] M. Yurt, S. U. H. Dar, B. Tınaz, M. Özbey, and T. Çukur, “Semi-supervised learning of mutually accelerated multi-contrast mri synthesis without fully-sampled ground-truths,” *arXiv preprint arXiv:2011.14347*, 2020.
- [64] M. Yurt, M. Özbey, S. U. H. Dar, B. Tınaz, K. K. Oğuz, and T. Çukur, “Progressively volumetrized deep generative models for data-efficient contextual learning of mr image recovery,” *arXiv preprint arXiv:2011.13913*, 2020.
- [65] Y. Korkmaz, S. U. Dar, M. Yurt, M. Özbey, and T. Çukur, “Unsupervised mri reconstruction via zero-shot learned adversarial transformers,” *arXiv preprint arXiv:2105.08059*, 2021.
- [66] S. U. Dar, M. Yurt, M. Shahdloo, M. E. Ildız, B. Tınaz, and T. Çukur, “Prior-guided image reconstruction for accelerated multi-contrast mri via generative adversarial networks,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1072–1087, 2020.
- [67] S. U. H. Dar, M. Özbey, A. B. Çatlı, and T. Çukur, “A transfer-learning approach for accelerated mri using deep neural networks,” *Magnetic resonance in medicine*, vol. 84, no. 2, pp. 663–685, 2020.
- [68] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, “Image synthesis in multi-contrast mri with conditional generative adversarial networks,” *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2375–2388, 2019.

- [69] K. T. Kwon, H. H. Wu, T. Shin, T. Çukur, M. Lustig, and D. G. Nishimura, “Three-dimensional magnetization-prepared imaging using a concentric cylinders trajectory,” *Magnetic resonance in medicine*, vol. 71, no. 5, pp. 1700–1710, 2014.
- [70] G. B. Chavhan, P. S. Babyn, B. Thomas, M. M. Shroff, and E. M. Haacke, “Principles, techniques, and applications of t2*-based mr imaging and its special applications,” *Radiographics*, vol. 29, no. 5, pp. 1433–1449, 2009.
- [71] L. Pauling and C. D. Coryell, “The magnetic properties and structure of hemoglobin, oxyhemoglobin and carbonmonoxyhemoglobin,” *Proceedings of the National Academy of Sciences*, vol. 22, no. 4, pp. 210–216, 1936.
- [72] P. T. Fox and M. E. Raichle, “Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects,” *Proceedings of the National Academy of Sciences*, vol. 83, no. 4, pp. 1140–1144, 1986.
- [73] P. T. Fox, M. E. Raichle, M. A. Mintun, and C. Dence, “Nonoxidative glucose consumption during focal physiologic neural activity,” *Science*, vol. 241, no. 4864, pp. 462–464, 1988.
- [74] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, “Neurophysiological investigation of the basis of the fmri signal,” *nature*, vol. 412, no. 6843, pp. 150–157, 2001.
- [75] R. A. Poldrack, J. A. Mumford, and T. E. Nichols, *Handbook of functional MRI data analysis*. Cambridge University Press, 2011.
- [76] G. K. Aguirre, E. Zarahn, and M. D’Esposito, “The variability of human, bold hemodynamic responses,” *Neuroimage*, vol. 8, no. 4, pp. 360–369, 1998.
- [77] T. Çukur, A. G. Huth, S. Nishimoto, and J. L. Gallant, “Functional subdomains within scene-selective cortex: parahippocampal place area, retrosplenial complex, and occipital place area,” *Journal of Neuroscience*, vol. 36, no. 40, pp. 10257–10273, 2016.

- [78] T. Çukur, A. G. Huth, S. Nishimoto, and J. L. Gallant, “Functional subdomains within human ffa,” *Journal of Neuroscience*, vol. 33, no. 42, pp. 16748–16766, 2013.
- [79] T. Çukur, S. Nishimoto, A. G. Huth, and J. L. Gallant, “Attention during natural vision warps semantic representation across the human brain,” *Nature neuroscience*, vol. 16, no. 6, pp. 763–770, 2013.
- [80] M. Aksoy, C. Forman, M. Straka, T. Çukur, J. Hornegger, and R. Bammer, “Hybrid prospective and retrospective head motion correction to mitigate cross-calibration errors,” *Magnetic resonance in medicine*, vol. 67, no. 5, pp. 1237–1251, 2012.
- [81] I. Kiremitçi, Ö. Yilmaz, E. Çelik, M. Shahdloo, A. G. Huth, and T. Çukur, “Attentional modulation of hierarchical speech representations in a multi-talker environment,” *Cerebral Cortex*, 2021.
- [82] M. Shahdloo, E. Çelik, B. A. Urgen, J. L. Gallant, and T. Çukur, “Task-dependent warping of semantic representations during search for visual action categories,” *bioRxiv*, 2021.
- [83] T. Zhang, J. S. Gao, T. Çukur, and J. L. Gallant, “Voxel-based state space modeling recovers task-related cognitive states in naturalistic fmri experiments,” *Frontiers in neuroscience*, vol. 14, 2020.
- [84] Ö. Yılmaz, E. Çelik, and T. Çukur, “Informed feature regularization in voxelwise modeling for naturalistic fmri experiments,” *The European journal of neuroscience*, vol. 52, no. 5, pp. 3394–3410, 2020.
- [85] E. Çelik, S. U. H. Dar, Ö. Yılmaz, Ü. Keleş, and T. Çukur, “Spatially informed voxelwise modeling for naturalistic fmri experiments,” *NeuroImage*, vol. 186, pp. 741–757, 2019.
- [86] N. Moray, “Attention in dichotic listening: Affective cues and the influence of instructions,” *Quarterly journal of experimental psychology*, vol. 11, no. 1, pp. 56–60, 1959.

- [87] A. M. Treisman, “Contextual cues in selective listening,” *Quarterly Journal of Experimental Psychology*, vol. 12, no. 4, pp. 242–248, 1960.
- [88] J. A. Deutsch and D. Deutsch, “Attention: Some theoretical considerations,” *Psychological review*, vol. 70, no. 1, p. 80, 1963.
- [89] R. Corteen and D. Dunn, “Shock-associated words in a nonattended message: A test for momentary awareness,” *Journal of Experimental Psychology*, vol. 102, no. 6, p. 1143, 1974.
- [90] J. Driver, “A selective review of selective attention research from the past century,” *British Journal of Psychology*, vol. 92, no. 1, pp. 53–78, 2001.
- [91] D. E. Broadbent, *Perception and communication*. Elsevier, 2013.
- [92] S. A. Hillyard, R. F. Hink, V. L. Schwent, and T. W. Picton, “Electrical signs of selective attention in the human brain,” *Science*, pp. 177–180, 1973.
- [93] K. Alho, S. V. Medvedev, S. V. Pakhomov, M. S. Roudas, M. Tervaniemi, K. Reinikainen, T. Zeffiro, and R. Näätänen, “Selective tuning of the left and right auditory cortices during spatially directed attention,” *Cognitive Brain Research*, vol. 7, no. 3, pp. 335–341, 1999.
- [94] L. Jäncke, T. Buchanan, K. Lutz, and N. Shah, “Focused and nonfocused attention in verbal and emotional dichotic listening: an fmri study,” *Brain and language*, vol. 78, no. 3, pp. 349–363, 2001.
- [95] L. Jäncke, K. Specht, J. N. Shah, and K. Hugdahl, “Focused attention in a simple dichotic listening task: an fmri experiment,” *Cognitive Brain Research*, vol. 16, no. 2, pp. 257–266, 2003.
- [96] B. Lipschutz, R. Kolinsky, P. Damhaut, D. Wikler, and S. Goldman, “Attention-dependent changes of activation and connectivity in dichotic listening,” *Neuroimage*, vol. 17, no. 2, pp. 643–656, 2002.
- [97] C. I. Petkov, X. Kang, K. Alho, O. Bertrand, E. W. Yund, and D. L. Woods, “Attentional modulation of human auditory cortex,” *Nature neuroscience*, vol. 7, no. 6, pp. 658–663, 2004.

- [98] J. A. Johnson and R. J. Zatorre, “Attention to simultaneous unrelated auditory and visual events: behavioral and neural correlates,” *Cerebral Cortex*, vol. 15, no. 10, pp. 1609–1620, 2005.
- [99] A. Degerman, T. Rinne, J. Salmi, O. Salonen, and K. Alho, “Selective attention to sound location or pitch studied with fmri,” *Brain research*, vol. 1077, no. 1, pp. 123–134, 2006.
- [100] T. Rinne, J. Pekkola, A. Degerman, T. Autti, I. P. Jääskeläinen, M. Sams, and K. Alho, “Modulation of auditory cortex activation by sound presentation rate and attention,” *Human brain mapping*, vol. 26, no. 2, pp. 94–99, 2005.
- [101] T. Rinne, M. H. Balk, S. Koistinen, T. Autti, K. Alho, and M. Sams, “Auditory selective attention modulates activation of human inferior colliculus,” *Journal of Neurophysiology*, vol. 100, no. 6, pp. 3323–3327, 2008.
- [102] T. Rinne, “Activations of human auditory cortex during visual and auditory selective attention tasks with varying difficulty,” *The open neuroimaging journal*, vol. 4, p. 187, 2010.
- [103] D. L. Woods, G. C. Stecker, T. Rinne, T. J. Herron, A. D. Cate, E. W. Yund, I. Liao, and X. Kang, “Functional maps of human auditory cortex: effects of acoustic features and attention,” *PloS one*, vol. 4, no. 4, p. e5183, 2009.
- [104] D. L. Woods, T. J. Herron, A. D. Cate, E. W. Yund, G. C. Stecker, T. Rinne, and X. Kang, “Functional properties of human auditory cortical fields,” *Frontiers in systems neuroscience*, vol. 4, p. 155, 2010.
- [105] A. E. Paltoglou, C. J. Sumner, and D. A. Hall, “Examining the role of frequency specificity in the enhancement and suppression of human cortical activity by auditory selective attention,” *Hearing research*, vol. 257, no. 1-2, pp. 106–118, 2009.
- [106] G. Kidd and H. S. Colburn, “Informational masking in speech recognition,” *The auditory system at the cocktail party*, pp. 75–109, 2017.

- [107] Y. Li, F. Wang, Y. Chen, A. Cichocki, and T. Sejnowski, “The effects of audiovisual inputs on solving the cocktail party problem in the human brain: An fmri study,” *Cerebral Cortex*, vol. 28, no. 10, pp. 3623–3637, 2018.
- [108] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, “Auditory attention—focusing the searchlight on sound,” *Current opinion in neurobiology*, vol. 17, no. 4, pp. 437–455, 2007.
- [109] A. W. Bronkhorst, “The cocktail-party problem revisited: early processing and selection of multi-talker speech,” *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, 2015.
- [110] R. F. Hink and S. A. Hillyard, “Auditory evoked potentials during selective listening to dichotic speech messages,” *Perception & Psychophysics*, vol. 20, no. 4, pp. 236–242, 1976.
- [111] W. Teder, T. Kujala, and R. Näätänen, “Selection of speech messages in free-field listening,” *Neuroreport: An International Journal for the Rapid Communication of Research in Neuroscience*, 1993.
- [112] K. Alho, V. A. Vorobyev, S. V. Medvedev, S. V. Pakhomov, M. S. Roudas, M. Tervaniemi, T. van Zuijen, and R. Näätänen, “Hemispheric lateralization of cerebral blood-flow changes during selective listening to dichotically presented continuous speech,” *Cognitive brain research*, vol. 17, no. 2, pp. 201–211, 2003.
- [113] M. Elhilali, J. Xiang, S. A. Shamma, and J. Z. Simon, “Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene,” *PLoS Biol*, vol. 7, no. 6, p. e1000129, 2009.
- [114] A. Gutschalk and A. R. Dykstra, “Functional imaging of auditory scene analysis,” *Hearing research*, vol. 307, pp. 98–110, 2014.
- [115] M. H. Davis and I. S. Johnsrude, “Hierarchical processing in spoken language comprehension,” *Journal of Neuroscience*, vol. 23, no. 8, pp. 3423–3431, 2003.

- [116] T. D. Griffiths and J. D. Warren, “What is an auditory object?,” *Nature Reviews Neuroscience*, vol. 5, no. 11, pp. 887–892, 2004.
- [117] K. Okada, F. Rong, J. Venezia, W. Matchin, I.-H. Hsieh, K. Saberi, J. T. Serences, and G. Hickok, “Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech,” *Cerebral Cortex*, vol. 20, no. 10, pp. 2486–2495, 2010.
- [118] G. M. Di Liberto, J. A. O’Sullivan, and E. C. Lalor, “Low-frequency cortical entrainment to speech reflects phoneme-level processing,” *Current Biology*, vol. 25, no. 19, pp. 2457–2465, 2015.
- [119] W. A. de Heer, A. G. Huth, T. L. Griffiths, J. L. Gallant, and F. E. Theunissen, “The hierarchical cortical organization of human speech processing,” *Journal of Neuroscience*, vol. 37, no. 27, pp. 6539–6557, 2017.
- [120] C. Brodbeck, A. Presacco, and J. Z. Simon, “Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension,” *NeuroImage*, vol. 172, pp. 162–174, 2018.
- [121] J. Z. Simon, “Human auditory neuroscience and the cocktail party problem,” in *The Auditory System at the Cocktail Party*, pp. 169–197, Springer, 2017.
- [122] J. R. Kerlin, A. J. Shahin, and L. M. Miller, “Attentional gain control of ongoing cortical speech representations in a “cocktail party”,” *Journal of Neuroscience*, vol. 30, no. 2, pp. 620–628, 2010.
- [123] N. Ding and J. Z. Simon, “Neural coding of continuous speech in auditory cortex during monaural and dichotic listening,” *Journal of neurophysiology*, vol. 107, no. 1, pp. 78–89, 2012.
- [124] N. Ding and J. Z. Simon, “Emergence of neural encoding of auditory objects while listening to competing speakers,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11854–11859, 2012.

- [125] N. Mesgarani and E. F. Chang, “Selective cortical representation of attended speaker in multi-talker speech perception,” *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.
- [126] A. J. Power, J. J. Foxe, E.-J. Forde, R. B. Reilly, and E. C. Lalor, “At what time is the cocktail party? a late locus of selective attention to natural speech,” *European Journal of Neuroscience*, vol. 35, no. 9, pp. 1497–1503, 2012.
- [127] E. M. Z. Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon, *et al.*, “Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”,” *Neuron*, vol. 77, no. 5, pp. 980–991, 2013.
- [128] K. C. Puvvada and J. Z. Simon, “Cortical representations of speech in a multitalker auditory scene,” *Journal of Neuroscience*, vol. 37, no. 38, pp. 9189–9196, 2017.
- [129] C. Brodbeck, L. E. Hong, and J. Z. Simon, “Rapid transformation from auditory to linguistic representations of continuous speech,” *Current Biology*, vol. 28, no. 24, pp. 3976–3983, 2018.
- [130] J. O’Sullivan, J. Herrero, E. Smith, C. Schevon, G. M. McKhann, S. A. Sheth, A. D. Mehta, and N. Mesgarani, “Hierarchical encoding of attended auditory objects in multi-talker speech perception,” *Neuron*, vol. 104, no. 6, pp. 1195–1209, 2019.
- [131] S. Puschmann, S. Baillet, and R. J. Zatorre, “Musicians at the cocktail party: neural substrates of musical training during selective listening in multispeaker situations,” *Cerebral Cortex*, vol. 29, no. 8, pp. 3253–3265, 2019.
- [132] T. Nakai, C. Kato, and K. Matsuo, “An fmri study to investigate auditory attention: a model of the cocktail party phenomenon,” *Magnetic Resonance in Medical Sciences*, vol. 4, no. 2, pp. 75–82, 2005.

- [133] K. Alho, V. A. Vorobyev, S. V. Medvedev, S. V. Pakhomov, M. G. Starchenko, M. Tervaniemi, and R. Näätänen, “Selective attention to human voice enhances brain activity bilaterally in the superior temporal sulcus,” *Brain research*, vol. 1075, no. 1, pp. 142–150, 2006.
- [134] Y. Ikeda, N. Yahata, H. Takahashi, M. Koeda, K. Asai, Y. Okubo, and H. Suzuki, “Cerebral activation associated with speech sound discrimination during the diotic listening task: An fmri study,” *Neuroscience research*, vol. 67, no. 1, pp. 65–71, 2010.
- [135] K. T. Hill and L. M. Miller, “Auditory attentional control and selection during cocktail party listening,” *Cerebral cortex*, vol. 20, no. 3, pp. 583–590, 2010.
- [136] C. J. Wild, A. Yusuf, D. E. Wilson, J. E. Peelle, M. H. Davis, and I. S. Johnsrude, “Effortful listening: the processing of degraded speech depends critically on attention,” *Journal of Neuroscience*, vol. 32, no. 40, pp. 14010–14021, 2012.
- [137] M. Regev, E. Simony, K. Lee, K. M. Tan, J. Chen, and U. Hasson, “Propagation of information along the cortical hierarchy as a function of attention while reading and listening to stories,” *Cerebral Cortex*, vol. 29, no. 10, pp. 4017–4034, 2019.
- [138] P. Wikman, E. Sahari, V. Salmela, A. Leminen, M. Leminen, M. Laine, and K. Alho, “Breaking down the cocktail party: Attentional modulation of cerebral audiovisual speech processing,” *NeuroImage*, vol. 224, p. 117365, 2021.
- [139] S. K. Scott, S. Rosen, L. Wickham, and R. J. Wise, “A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception,” *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 813–821, 2004.
- [140] S. K. Scott, S. Rosen, C. P. Beaman, J. P. Davis, and R. J. Wise, “The neural processing of masked speech: Evidence for different mechanisms in

- the left and right temporal lobes,” *The Journal of the Acoustical Society of America*, vol. 125, no. 3, pp. 1737–1743, 2009.
- [141] S. K. Scott and C. McGettigan, “The neural processing of masked speech,” *Hearing research*, vol. 303, pp. 58–66, 2013.
 - [142] S. Evans, C. McGettigan, Z. K. Agnew, S. Rosen, and S. K. Scott, “Getting the cocktail party started: masking effects in speech perception,” *Journal of cognitive neuroscience*, vol. 28, no. 3, pp. 483–500, 2016.
 - [143] A. G. Huth, W. A. De Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, “Natural speech reveals the semantic maps that tile human cerebral cortex,” *Nature*, vol. 532, no. 7600, pp. 453–458, 2016.
 - [144] M. Jenkinson and S. Smith, “A global optimisation method for robust affine registration of brain images,” *Medical image analysis*, vol. 5, no. 2, pp. 143–156, 2001.
 - [145] S. M. Smith, “Fast robust automated brain extraction,” *Human brain mapping*, vol. 17, no. 3, pp. 143–155, 2002.
 - [146] J. S. Gao, A. G. Huth, M. D. Lescroart, and J. L. Gallant, “Pycortex: an interactive surface visualizer for fmri,” *Frontiers in neuroinformatics*, vol. 9, p. 23, 2015.
 - [147] A. M. Dale, B. Fischl, and M. I. Sereno, “Cortical surface-based analysis: I. segmentation and surface reconstruction,” *Neuroimage*, vol. 9, no. 2, pp. 179–194, 1999.
 - [148] C. Destrieux, B. Fischl, A. Dale, and E. Halgren, “Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature,” *Neuroimage*, vol. 53, no. 1, pp. 1–15, 2010.
 - [149] D. N. Greve and B. Fischl, “Accurate and robust brain image alignment using boundary-based registration,” *Neuroimage*, vol. 48, no. 1, pp. 63–72, 2009.

- [150] S. Da Costa, W. van der Zwaag, J. P. Marques, R. S. Frackowiak, S. Clarke, and M. Saenz, “Human primary auditory cortex follows the shape of heschl’s gyrus,” *Journal of Neuroscience*, vol. 31, no. 40, pp. 14067–14075, 2011.
- [151] R. Lyon, “A computational model of filtering, detection, and compression in the cochlea,” in *ICASSP’82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, pp. 1282–1285, IEEE, 1982.
- [152] M. Slaney, “Auditory toolbox,” *Interval Research Corporation, Tech. Rep*, vol. 10, no. 1998, p. 1194, 1998.
- [153] P. Gill, J. Zhang, S. M. Woolley, T. Fremouw, and F. E. Theunissen, “Sound representation methods for spectro-temporal receptive field estimation,” *Journal of computational neuroscience*, vol. 21, no. 1, p. 5, 2006.
- [154] W. J. Levelt, *Speaking: From intention to articulation*, vol. 1. MIT press, 1993.
- [155] J. Yuan and M. Liberman, “Speaker identification on the scotus corpus,” *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.
- [156] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, “Predicting human brain activity associated with the meanings of nouns,” *science*, vol. 320, no. 5880, pp. 1191–1195, 2008.
- [157] I. DeWitt and J. P. Rauschecker, “Phoneme and word recognition in the auditory ventral stream,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 8, pp. E505–E514, 2012.
- [158] J. K. Bizley and Y. E. Cohen, “The what, where and how of auditory-object perception,” *Nature Reviews Neuroscience*, vol. 14, no. 10, pp. 693–707, 2013.
- [159] S. K. Scott, “Auditory processing—speech, space and auditory objects,” *Current opinion in neurobiology*, vol. 15, no. 2, pp. 197–201, 2005.

- [160] J. P. Rauschecker and B. Tian, “Mechanisms and streams for processing of “what” and “where” in auditory cortex,” *Proceedings of the National Academy of Sciences*, vol. 97, no. 22, pp. 11800–11806, 2000.
- [161] D. Saur, B. W. Kreher, S. Schnell, D. Kümmerer, P. Kellmeyer, M.-S. Vry, R. Umarova, M. Musso, V. Glauche, S. Abel, *et al.*, “Ventral and dorsal pathways for language,” *Proceedings of the national academy of Sciences*, vol. 105, no. 46, pp. 18035–18040, 2008.
- [162] J. P. Rauschecker, “An expanded role for the dorsal auditory pathway in sensorimotor control and integration,” *Hearing research*, vol. 271, no. 1-2, pp. 16–25, 2011.
- [163] A. D. Friederici, “The cortical language circuit: from auditory perception to sentence comprehension,” *Trends in cognitive sciences*, vol. 16, no. 5, pp. 262–268, 2012.
- [164] T. Ueno, S. Saito, T. T. Rogers, and M. A. L. Ralph, “Lichtheim 2: synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways,” *Neuron*, vol. 72, no. 2, pp. 385–396, 2011.
- [165] I. Bornkessel-Schlesewsky and M. Schlewsky, “Reconciling time, space and function: a new dorsal–ventral stream model of sentence comprehension,” *Brain and language*, vol. 125, no. 1, pp. 60–76, 2013.
- [166] I. Bornkessel-Schlesewsky, M. Schlewsky, S. L. Small, and J. P. Rauschecker, “Neurobiological roots of language in primate audition: common computational properties,” *Trends in cognitive sciences*, vol. 19, no. 3, pp. 142–150, 2015.
- [167] B. Bernal and A. Ardila, “The role of the arcuate fasciculus in conduction aphasia,” *Brain*, vol. 132, no. 9, pp. 2309–2316, 2009.
- [168] G. Hickok and D. Poeppel, “Towards a functional neuroanatomy of speech perception,” *Trends in cognitive sciences*, vol. 4, no. 4, pp. 131–138, 2000.

- [169] S. Frey, J. S. Campbell, G. B. Pike, and M. Petrides, “Dissociating the human language pathways with high angular resolution diffusion fiber tractography,” *Journal of Neuroscience*, vol. 28, no. 45, pp. 11435–11444, 2008.
- [170] S. K. Scott, C. McGettigan, and F. Eisner, “A little more conversation, a little less action—candidate roles for the motor cortex in speech perception,” *Nature Reviews Neuroscience*, vol. 10, no. 4, pp. 295–302, 2009.
- [171] R. J. Zatorre and P. Belin, “Spectral and temporal processing in human auditory cortex,” *Cerebral cortex*, vol. 11, no. 10, pp. 946–953, 2001.
- [172] J. Obleser, F. Eisner, and S. A. Kotz, “Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features,” *Journal of neuroscience*, vol. 28, no. 32, pp. 8116–8123, 2008.
- [173] C. McGettigan and S. K. Scott, “Cortical asymmetries in speech perception: what’s wrong, what’s right and what’s left?,” *Trends in cognitive sciences*, vol. 16, no. 5, pp. 269–276, 2012.
- [174] M. Corbetta, G. Patel, and G. L. Shulman, “The reorienting system of the human brain: from environment to theory of mind,” *Neuron*, vol. 58, no. 3, pp. 306–324, 2008.
- [175] S. Vossel, J. J. Geng, and G. R. Fink, “Dorsal and ventral attention systems: distinct neural circuits but collaborative roles,” *The Neuroscientist*, vol. 20, no. 2, pp. 150–159, 2014.
- [176] S. Puschmann, S. Steinkamp, I. Gillich, B. Mirkovic, S. Debener, and C. M. Thiel, “The right temporoparietal junction supports speech tracking during selective listening: Evidence from concurrent eeg-fmri,” *Journal of Neuroscience*, vol. 37, no. 47, pp. 11505–11516, 2017.
- [177] J. R. Binder, J. A. Frost, T. A. Hammeke, P. S. Bellgowan, J. A. Springer, J. N. Kaufman, and E. T. Possing, “Human temporal lobe activation by speech and nonspeech sounds,” *Cerebral cortex*, vol. 10, no. 5, pp. 512–528, 2000.

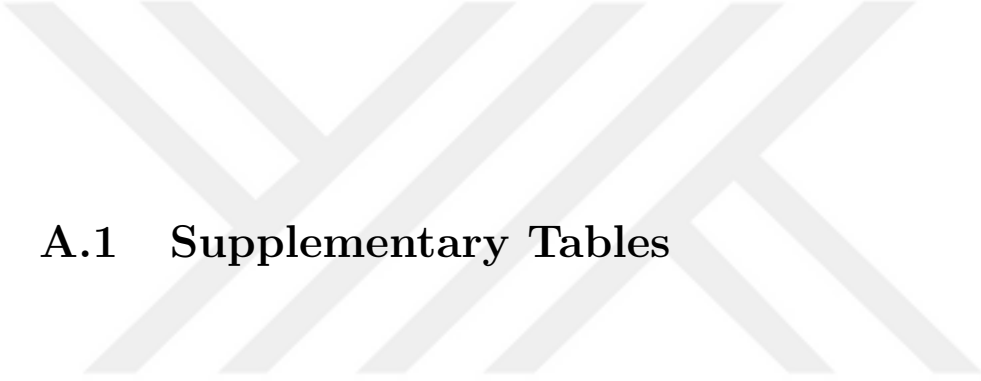
- [178] E. Liebenthal, J. R. Binder, S. M. Spitzer, E. T. Possing, and D. A. Medler, “Neural substrates of phonemic perception,” *Cerebral cortex*, vol. 15, no. 10, pp. 1621–1631, 2005.
- [179] J. Obleser, J. Zimmermann, J. Van Meter, and J. P. Rauschecker, “Multiple stages of auditory speech perception reflected in event-related fmri,” *Cerebral Cortex*, vol. 17, no. 10, pp. 2251–2257, 2007.
- [180] S. Da Costa, W. van der Zwaag, L. M. Miller, S. Clarke, and M. Saenz, “Tuning in to sound: frequency-selective attentional filter in human primary auditory cortex,” *Journal of Neuroscience*, vol. 33, no. 5, pp. 1858–1863, 2013.
- [181] A. Seydell-Greenwald, A. S. Greenberg, and J. P. Rauschecker, “Are you listening? brain activation associated with sustained nonspatial auditory attention in the presence and absence of stimulation,” *Human brain mapping*, vol. 35, no. 5, pp. 2233–2252, 2014.
- [182] L. Riecke, J. C. Peters, G. Valente, V. G. Kemper, E. Formisano, and B. Sorger, “Frequency-selective attention in auditory scenes recruits frequency representations throughout human superior temporal cortex,” *Cerebral Cortex*, vol. 27, no. 5, pp. 3002–3014, 2017.
- [183] N. Ding and J. Z. Simon, “Cortical entrainment to continuous speech: functional roles and interpretations,” *Frontiers in human neuroscience*, vol. 8, p. 311, 2014.
- [184] M. Corbetta and G. L. Shulman, “Control of goal-directed and stimulus-driven attention in the brain,” *Nature reviews neuroscience*, vol. 3, no. 3, pp. 201–215, 2002.
- [185] A. M. Liberman and I. G. Mattingly, “The motor theory of speech perception revised,” *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.
- [186] M. H. Davis and I. S. Johnsrude, “Hearing speech sounds: top-down influences on the interface between audition and speech perception,” *Hearing research*, vol. 229, no. 1-2, pp. 132–147, 2007.

- [187] R. Möttönen, R. Dutton, and K. E. Watkins, “Auditory-motor processing of speech sounds,” *Cerebral Cortex*, vol. 23, no. 5, pp. 1190–1197, 2013.
- [188] B. Osnes, K. Hugdahl, and K. Specht, “Effective connectivity analysis demonstrates involvement of premotor cortex during speech perception,” *Neuroimage*, vol. 54, no. 3, pp. 2437–2445, 2011.
- [189] A. G. Hervais-Adelman, R. P. Carlyon, I. S. Johnsrude, and M. H. Davis, “Brain regions recruited for the effortful comprehension of noise-vocoded words,” *Language and Cognitive Processes*, vol. 27, no. 7-8, pp. 1145–1166, 2012.
- [190] S. Rutten, R. Santoro, A. Hervais-Adelman, E. Formisano, and N. Golestani, “Cortical encoding of speech enhances task-relevant acoustic information,” *Nature human behaviour*, vol. 3, no. 9, pp. 974–987, 2019.
- [191] N. Lavie, “Distracted and confused?: Selective attention under load,” *Trends in cognitive sciences*, vol. 9, no. 2, pp. 75–82, 2005.
- [192] V. Boulenger, M. Hoen, E. Ferragne, F. Pellegrino, and F. Meunier, “Real-time lexical competitions during speech-in-speech comprehension,” *Speech Communication*, vol. 52, no. 3, pp. 246–253, 2010.
- [193] M. Sabri, J. R. Binder, R. Desai, D. A. Medler, M. D. Leitl, and E. Liebenthal, “Attentional and linguistic interactions in speech perception,” *Neuroimage*, vol. 39, no. 3, pp. 1444–1456, 2008.
- [194] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.
- [195] F. Pulvermüller and Y. Shtyrov, “Language outside the focus of attention: the mismatch negativity as a tool for studying higher cognitive processes,” *Progress in neurobiology*, vol. 79, no. 1, pp. 49–71, 2006.
- [196] K. Relander, P. Rämä, and T. Kujala, “Word semantics is processed even without attentional effort,” *Journal of Cognitive Neuroscience*, vol. 21, no. 8, pp. 1511–1522, 2009.

- [197] R. Näätänen, T. Kujala, and I. Winkler, “Auditory processing that leads to conscious perception: a unique window to central auditory processing opened by the mismatch negativity and related responses,” *Psychophysiology*, vol. 48, no. 1, pp. 4–22, 2011.
- [198] P. Rämä, K. Relander-Syrjänen, S. Carlson, O. Salonen, and T. Kujala, “Attention and semantic processing during speech: an fmri study,” *Brain and language*, vol. 122, no. 2, pp. 114–119, 2012.
- [199] N. Ding, X. Pan, C. Luo, N. Su, W. Zhang, and J. Zhang, “Attention is required for knowledge-based sequential grouping: insights from the integration of syllables into words,” *Journal of Neuroscience*, vol. 38, no. 5, pp. 1178–1188, 2018.
- [200] A. Treisman, “Monitoring and storage of irrelevant messages in selective attention,” *Journal of Verbal Learning and Verbal Behavior*, vol. 3, no. 6, pp. 449–459, 1964.
- [201] N. Wood and N. Cowan, “The cocktail party phenomenon revisited: how frequent are attention shifts to one’s name in an irrelevant auditory channel?,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 21, no. 1, p. 255, 1995.
- [202] J. L. Lewis, “Semantic processing of unattended messages using dichotic listening,” *Journal of experimental psychology*, vol. 85, no. 2, p. 225, 1970.
- [203] M. Rivenez, C. J. Darwin, and A. Guillaume, “Processing unattended speech,” *The journal of the Acoustical Society of America*, vol. 119, no. 6, pp. 4027–4040, 2006.
- [204] C. Spence and J. Driver, “Audiovisual links in exogenous covert spatial orienting,” *Perception & psychophysics*, vol. 59, no. 1, pp. 1–22, 1997.
- [205] F. B. Parmentier, J. Turner, and L. Perez, “A dual contribution to the involuntary semantic processing of unexpected spoken words,” *Journal of Experimental Psychology: General*, vol. 143, no. 1, p. 38, 2014.

Appendix A

Supplementary Materials



A.1 Supplementary Tables

Table A.1: Inter-run head motion in the cocktail-party experiment

		Rx	Ry	Rz	Tx	Ty	Tz
S1	Run2	0.0070	-0.0222	-0.0047	-0.8205	-1.8318	4.4034
	Run3	0.0043	-0.0193	-0.0075	0.3561	-1.1648	3.6750
	Run4	0.0081	-0.0400	-0.0122	0.2514	-2.3087	5.9718
	Run5	0.0059	-0.0258	-0.0040	-1.2676	-1.2530	4.4150
	Run6	0.0060	-0.0225	-0.0058	-0.5274	-1.9113	4.5966
S2	Run2	-0.0099	-0.0067	-0.0099	0.6676	-0.2590	-0.5229
	Run3	-0.0088	-0.0026	0.0018	-0.5063	0.6162	-0.6334
	Run4	-0.0128	-0.0118	-0.0103	-0.2114	0.3525	-0.3772
	Run5	-0.0119	-0.0044	-0.0047	-0.2807	0.7589	-1.1399
	Run6	-0.0101	-0.0066	-0.0118	0.7884	-0.1214	-0.4955
S3	Run2	0.0321	0.0118	0.0060	-0.3023	-0.1689	3.3696
	Run3	0.0280	0.0160	0.0059	-0.1685	-0.2995	1.9887
	Run4	0.0273	0.0139	0.0092	-0.7141	0.5719	2.3134
	Run5	0.0252	0.0129	-0.0008	0.3123	-0.5163	2.1007
	Run6	0.0263	0.0135	0.0055	-0.2489	-0.1244	2.4375
S4	Run2	-0.0009	0.0062	-0.0028	0.7449	-0.1245	-0.5775
	Run3	-0.0014	0.0042	0.0005	0.4494	0.5528	-0.5016
	Run4	0.0101	-0.0076	-0.0030	-0.0453	-0.7291	3.1879
	Run5	-0.0006	0.0019	-0.0051	1.0548	-0.5008	-0.0154
	Run6	-0.0010	0.0032	-0.0051	1.2785	-0.3303	-0.3824
S5	Run2	0.0106	-0.0133	-0.0164	3.3092	-1.8717	3.4716
	Run3	0.0181	-0.0245	-0.0150	3.8181	-1.6553	5.9240
	Run4	0.0097	-0.0330	-0.0067	2.0298	-0.6607	5.2784
	Run5	0.0156	-0.0299	-0.0011	0.7572	-0.5795	5.8197
	Run6	0.0079	-0.0250	-0.0202	5.3338	-1.7013	3.8040

Absolute motion between temporal-mean volume of each run (Run2-Run6) and temporal-mean volume of the first run (Run1) is calculated during the cocktail-party experiment. Motion estimates are obtained in subjects S1-S5 using FMRIB's Linear Image Registration Tool (FLIRT) (Jenkinson and Smith 2001). Rotational motion (Rx, Ry, Rz) and translational motion (Tx, Ty, Tz) are reported along the X, Y and Z axes. Rotations are expressed in radians, and translations are expressed in mm.

Table A.2: Number of significantly predicted voxels within ROIs in the left hemisphere

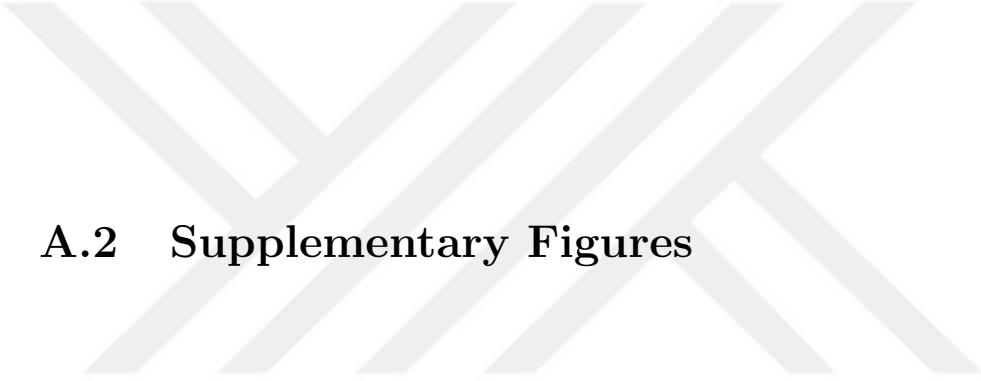
	Spectrally-sel.		Articulatorily-sel.		Semantically-sel.		Speech-sel.	
HG/HS _L	59.0	(±12.9)	63.9	(±8.3)	0.0	(±0.0)	101.5	(±15.0)
PT _L	29.7	(±7.4)	45.5	(±6.4)	20.0	(±3.5)	66.3	(±7.7)
pSF _L	22.0	(±5.5)	34.0	(±5.7)	6.1	(±3.8)	49.2	(±6.5)
STG _L	32.3	(±8.6)	163.8	(±27.6)	100.7	(±30.1)	280.2	(±40.2)
aSTG _L	6.4	(±3.4)	26.6	(±8.6)	42.4	(±9.8)	63.4	(±9.5)
mSTG _L	7.6	(±3.0)	82.5	(±14.0)	14.0	(±7.7)	92.7	(±18.3)
pSTG _L	12.3	(±4.9)	41.8	(±8.2)	30.4	(±9.9)	66.4	(±13.5)
STS _L	30.9	(±15.7)	200.2	(±22.8)	330.4	(±55.3)	426.7	(±59.1)
aSTS _L	10.2	(±3.9)	45.6	(±9.8)	75.0	(±12.3)	97.1	(±10.8)
mSTS _L	12.0	(±6.5)	70.8	(±11.4)	87.6	(±16.6)	123.1	(±17.2)
pSTS _L	12.3	(±8.1)	88.6	(±21.3)	165.6	(±22.7)	205.7	(±30.4)
MTG _L	7.1	(±4.8)	38.5	(±7.2)	137.3	(±41.3)	160.6	(±34.3)
SMG _L	24.9	(±7.70)	33.5	(±6.4)	72.7	(±22.9)	111.6	(±20.4)
AG _L	2.1	(±1.9)	4.2	(±2.4)	156.6	(±39.1)	157.6	(±39.7)
IPS _L	0.0	(±0.00)	0.0	(±0.0)	65.0	(±10.9)	65.0	(±10.9)
SPS _L	2.2	(±1.8)	5.8	(±1.5)	44.8	(±11.3)	49.2	(±10.6)
PrC _L	0.0	(±0.00)	9.1	(±5.1)	126.9	(±30.8)	129.9	(±31.7)
PCC _L	0.0	(±0.00)	0.0	(±0.0)	34.1	(±6.8)	34.1	(±6.8)
POS _L	0.0	(±0.00)	13.3	(±8.4)	58.7	(±18.2)	66.2	(±20.0)
POP _L	15.0	(±8.5)	24.6	(±6.7)	42.7	(±10.1)	64.8	(±9.9)
PTR _L	13.8	(±5.2)	31.8	(±7.5)	74.7	(±17.4)	98.3	(±17.2)
IFS _L	9.4	(±6.9)	26.3	(±9.1)	109.1	(±28.4)	124.4	(±30.0)
MFG _L	12.5	(±6.5)	38.2	(±10.5)	183.2	(±60.0)	216.1	(±56.4)
MFS _L	10.6	(±5.9)	25.1	(±9.1)	30.5	(±10.9)	57.6	(±8.3)
SFS _L	0.0	(±0.00)	2.4	(±2.1)	110.7	(±29.1)	111.5	(±29.1)
SFG _L	44.9	(±19.8)	59.9	(±12.7)	286.7	(±59.9)	327.6	(±62.6)
PreG _L	7.4	(±2.2)	41.9	(±7.5)	26.5	(±8.4)	60.4	(±12.2)
mOTS _L	0.0	(±0.0)	0.0	(±0.00)	26.3	(±3.1)	26.3	(±3.1)

Significantly predicted voxels by spectral, articulatory and semantic models are listed ($q(FDR) < 10^{-5}$, $t - test$; mean $\pm$ sem across subjects). The union of these voxels are taken as speech-selective voxels. Only ROIs in the left hemisphere and with at least 10 speech-selective voxels in each individual subject are listed. Subscripted “L” indicates right hemisphere.

Table A.3: Number of significantly predicted voxels within ROIs in the right hemisphere

	Spectrally-sel.		Articulatorily-sel.		Semantically-sel.		Speech-sel.	
HG/HS _R	47.4	(±5.6)	45.5	(±7.8)	0.0	(±0.0)	77.1	(±6.4)
PT _R	27.0	(±4.2)	29.6	(±3.7)	2.2	(±1.8)	41.7	(±3.4)
STG _R	56.9	(±8.5)	165.8	(±26.8)	55.1	(±8.7)	201.5	(±21.2)
aSTG _R	8.6	(±4.1)	30.5	(±11.8)	19.7	(±7.5)	47.4	(±12.6)
mSTG _R	12.3	(±4.0)	52.7	(±11.7)	7.0	(±2.0)	61.4	(±9.0)
pSTG _R	21.4	(±4.2)	57.2	(±11.0)	14.4	(±1.9)	71.6	(±10.2)
STS _R	39.9	(±12.1)	140.1	(±23.1)	338.3	(±45.6)	423.4	(±42.8)
aSTS _R	4.7	(±2.2)	30.0	(±8.0)	55.6	(±12.0)	71.1	(±13.1)
mSTS _R	24.6	(±8.9)	88.5	(±15.4)	86.1	(±24.1)	140.4	(±23.3)
pSTS _R	10.8	(±2.8)	25.0	(±4.6)	209.4	(±13.8)	224.2	(±12.5)
MTG _R	16.7	(±8.3)	29.5	(±9.1)	93.5	(±20.0)	114.7	(±20.3)
SMG _R	21.0	(±4.0)	28.2	(±6.8)	34.4	(±14.0)	71.0	(±13.7)
AG _R	2.1	(±1.8)	13.4	(±7.5)	137.1	(±35.6)	146.9	(±39.4)
IPS _R	0.0	(±0.0)	0.0	(±0.0)	33.4	(±4.0)	33.4	(±4.0)
SPS _R	2.0	(±1.7)	8.3	(±3.3)	66.0	(±17.1)	68.6	(±17.7)
PtC _R	0.0	(±0.0)	15.0	(±8.1)	107.7	(±29.9)	113.9	(±33.6)
PCC _R	0.0	(±0.0)	2.0	(±1.8)	36.2	(±8.4)	37.2	(±8.0)
POS _R	0.0	(±0.0)	18.3	(±9.5)	59.1	(±10.0)	73.5	(±15.5)
PTR _R	7.4	(±3.2)	18.3	(±6.1)	68.5	(±7.5)	80.7	(±7.2)
IFS _R	23.8	(±2.7)	41.5	(±12.0)	100.3	(±20.0)	138.3	(±21.6)
MFG _R	15.6	(±2.7)	51.1	(±15.3)	162.5	(±58.4)	218.2	(±50.6)
MFS _R	11.8	(±8.4)	29.4	(±9.7)	33.7	(±8.3)	70.5	(±11.4)
SFS _R	0.0	(±0.0)	12.6	(±5.5)	102.5	(±21.9)	109.3	(±21.6)
SFG _R	38.6	(±11.2)	75.4	(±21.7)	230.1	(±58.6)	295.3	(±60.5)

Significantly predicted voxels by spectral, articulatory and semantic models are listed ($q(FDR) < 10^{-5}$, $t - test$; mean $\pm$ sem across subjects). The union of these voxels are taken as speech-selective voxels. Only ROIs in the right hemisphere and with at least 10 speech-selective voxels in each individual subject are listed. Subscripted “R” indicates right hemisphere.



A.2 Supplementary Figures

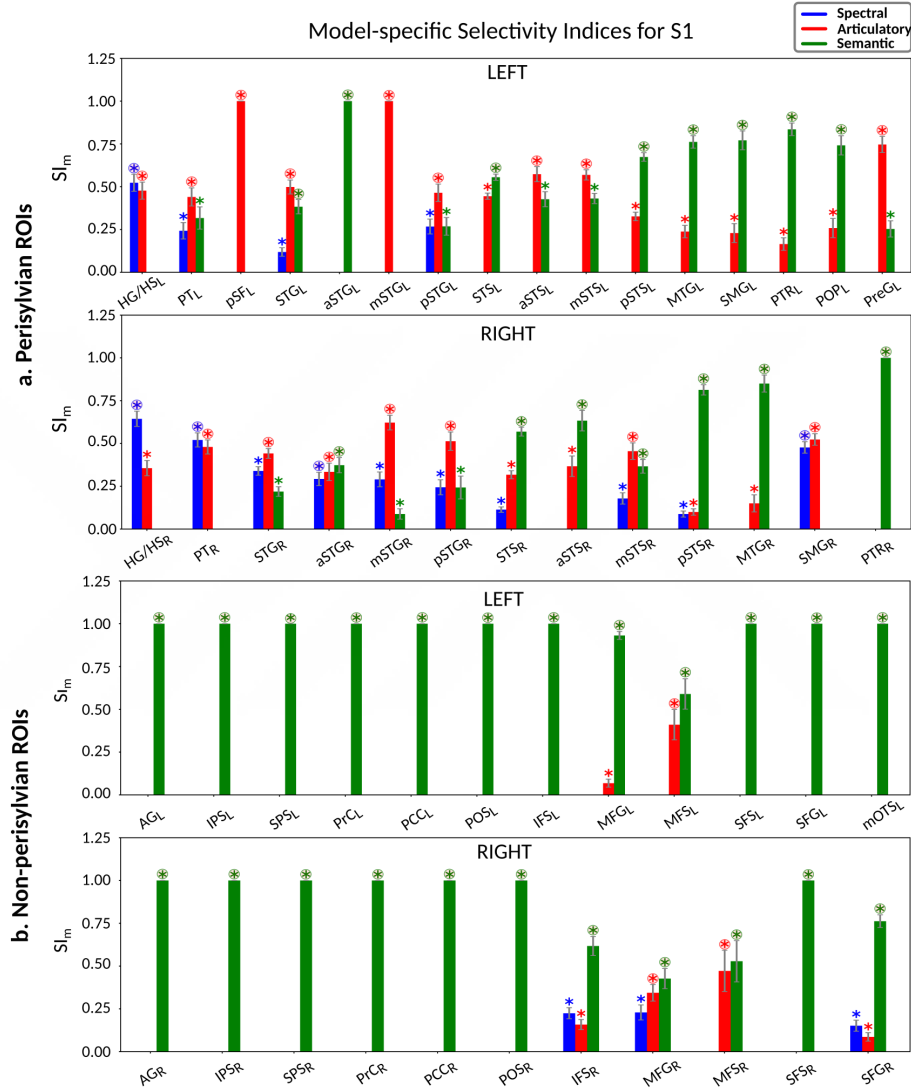


Figure A.1: **Model-specific selectivity indices for subject S1.** Single-voxel prediction scores on passive-listening data were used to quantify the degree of selectivity of each ROI to the underlying model features under passive listening. Prediction scores for a given model were averaged across speech-selective voxels within each ROI, and then normalized such that the cumulative prediction score from all models summed to 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory and semantic models (mean $\pm$ sem across ROI's speech-selective voxels). Significant indices are marked with *, colored according to the model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a. Perisylvian ROIs.** ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively.

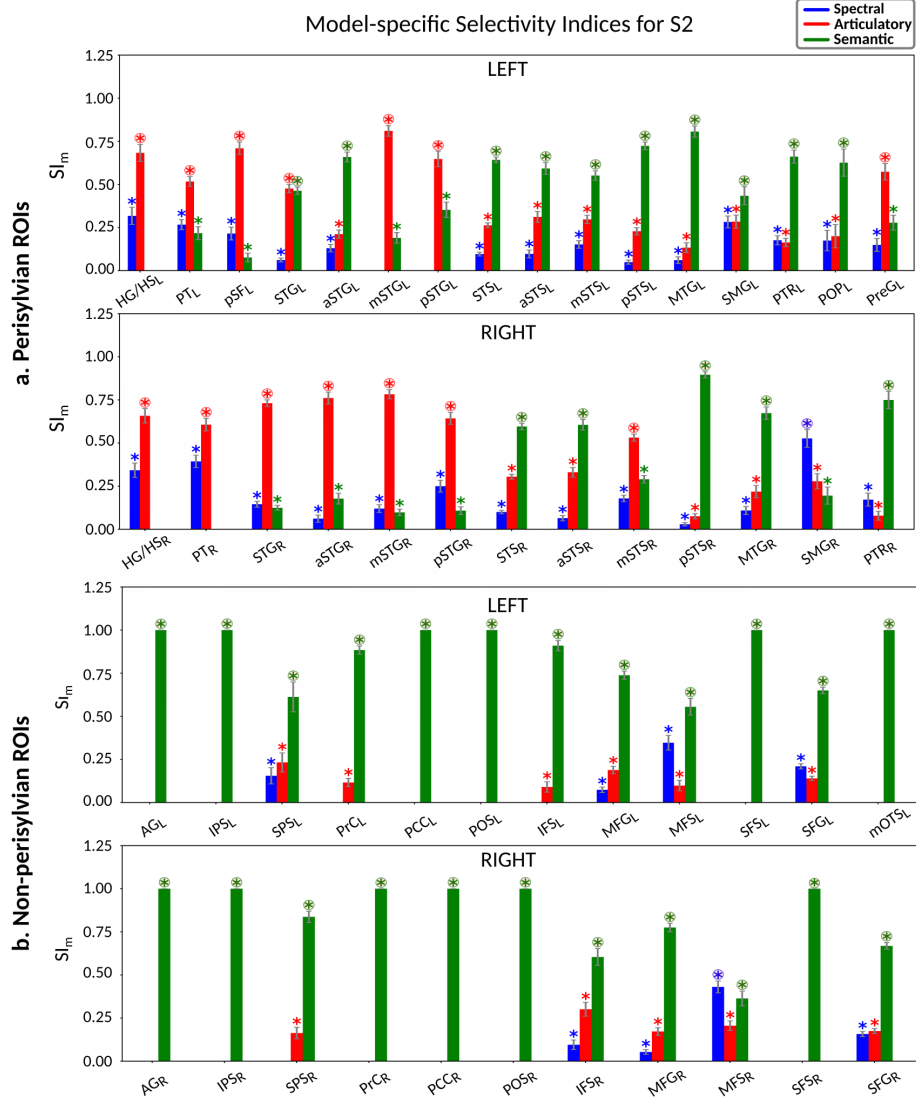


Figure A.2: **Model-specific selectivity indices for subject S2.** Single-voxel prediction scores on passive-listening data were used to quantify the degree of selectivity of each ROI to the underlying model features under passive listening. Prediction scores for a given model were averaged across speech-selective voxels within each ROI, and then normalized such that the cumulative prediction score from all models summed to 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory and semantic models (mean $\pm$ sem across ROI's speech-selective voxels). Significant indices are marked with *, colored according to the model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a. Perisylvian ROIs.** ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively.

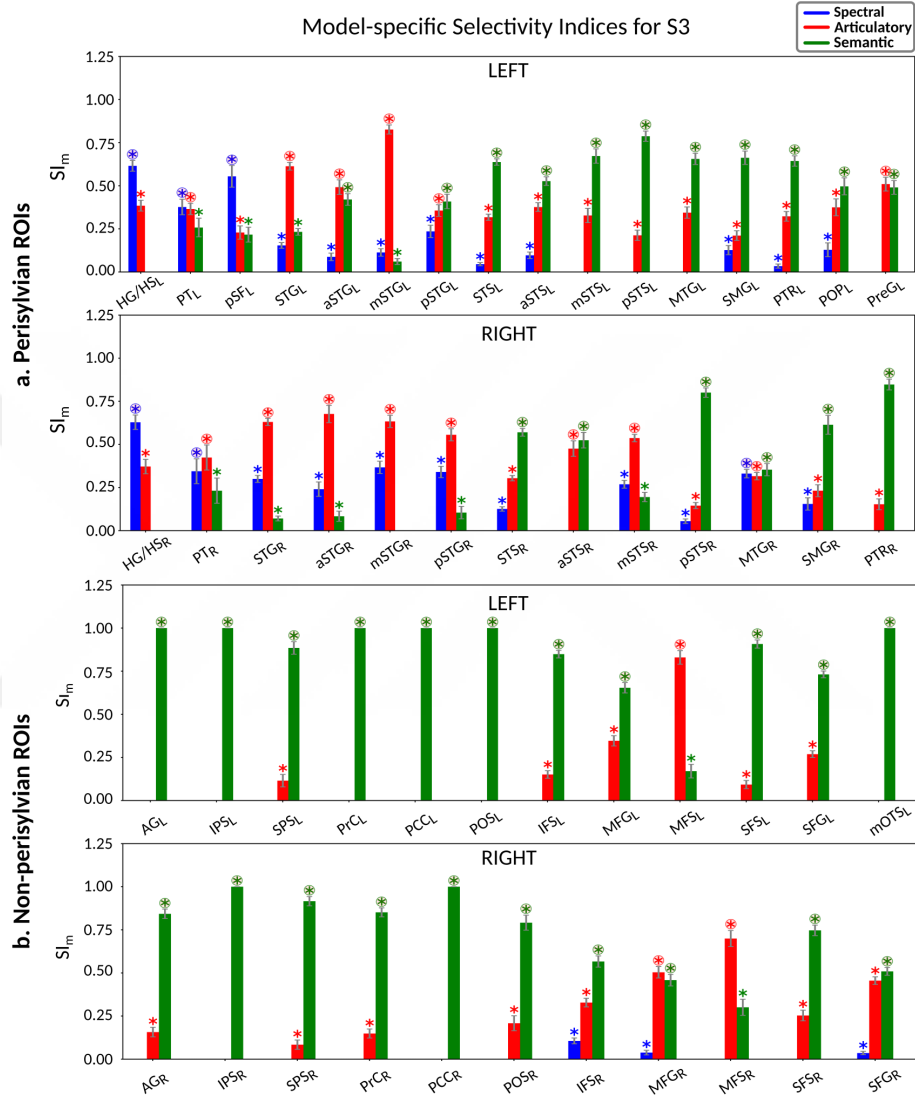


Figure A.3: **Model-specific selectivity indices for subject S3.** Single-voxel prediction scores on passive-listening data were used to quantify the degree of selectivity of each ROI to the underlying model features under passive listening. Prediction scores for a given model were averaged across speech-selective voxels within each ROI, and then normalized such that the cumulative prediction score from all models summed to 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory and semantic models (mean $\pm$ sem across ROI's speech-selective voxels). Significant indices are marked with *, colored according to the model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a. Perisylvian ROIs.** ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively.

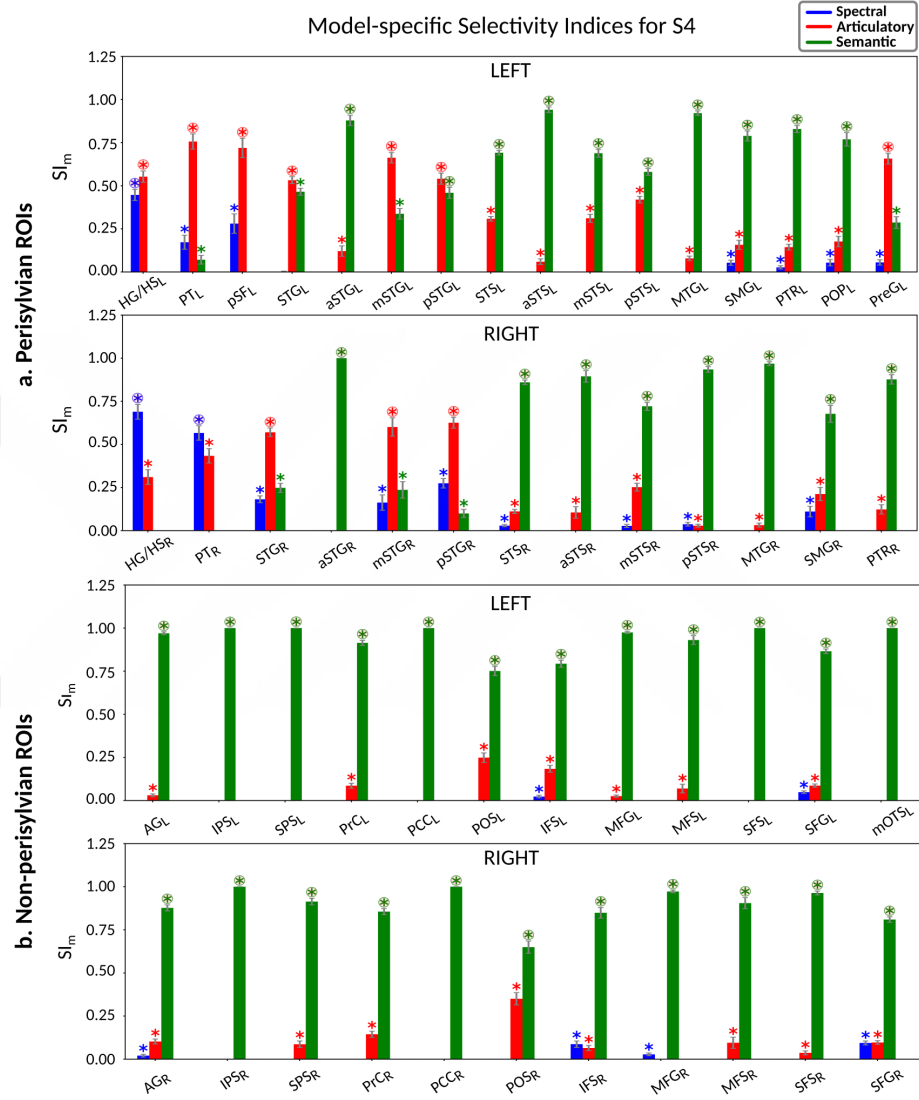


Figure A.4: **Model-specific selectivity indices for subject S4.** Single-voxel prediction scores on passive-listening data were used to quantify the degree of selectivity of each ROI to the underlying model features under passive listening. Prediction scores for a given model were averaged across speech-selective voxels within each ROI, and then normalized such that the cumulative prediction score from all models summed to 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory and semantic models (mean $\pm$ sem across ROI's speech-selective voxels). Significant indices are marked with *, colored according to the model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a. Perisylvian ROIs.** ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively.

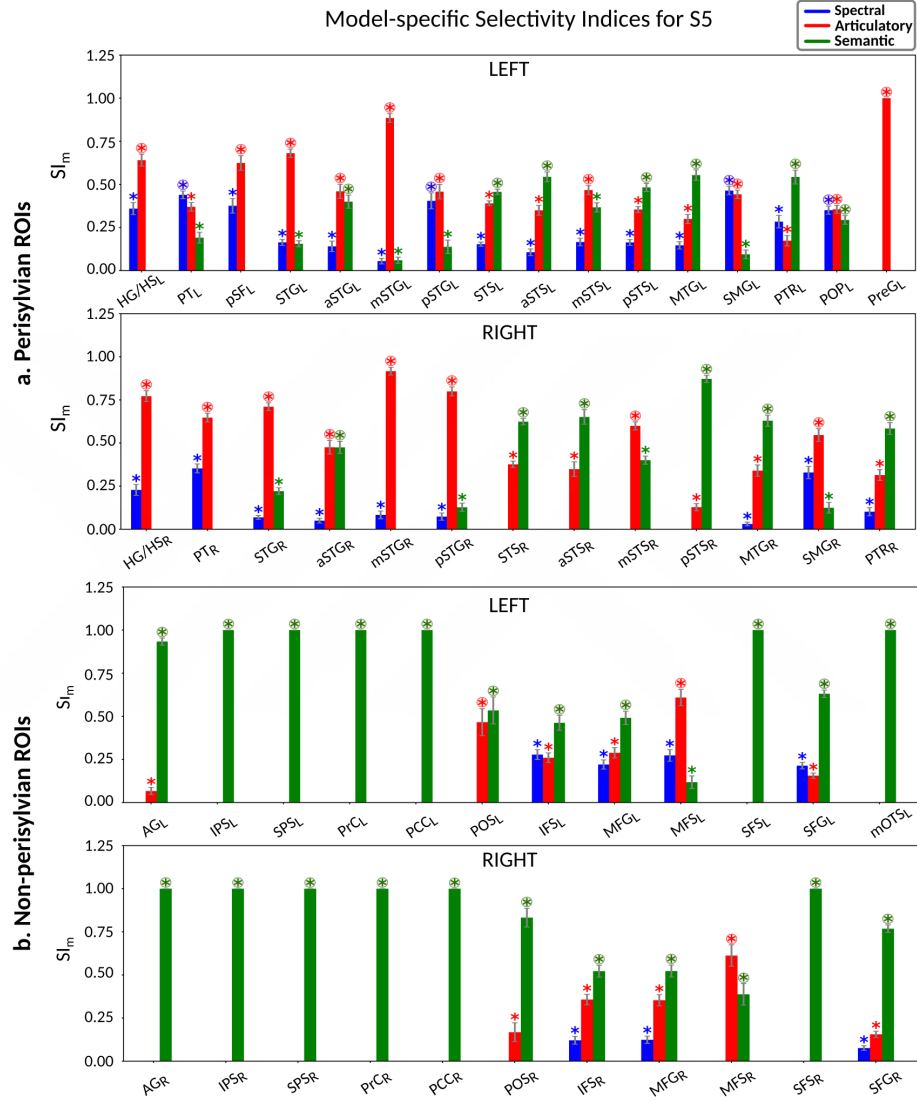


Figure A.5: **Model-specific selectivity indices for subject S5.** Single-voxel prediction scores on passive-listening data were used to quantify the degree of selectivity of each ROI to the underlying model features under passive listening. Prediction scores for a given model were averaged across speech-selective voxels within each ROI, and then normalized such that the cumulative prediction score from all models summed to 1. The resultant measure was taken as a model-specific selectivity index, (SI_m). SI_m is in the range of $[0, 1]$, where higher values indicate stronger selectivity for the underlying model. Bar plots display SI_m for spectral, articulatory and semantic models (mean $\pm$ sem across ROI's speech-selective voxels). Significant indices are marked with *, colored according to the model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a. Perisylvian ROIs.** ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively.

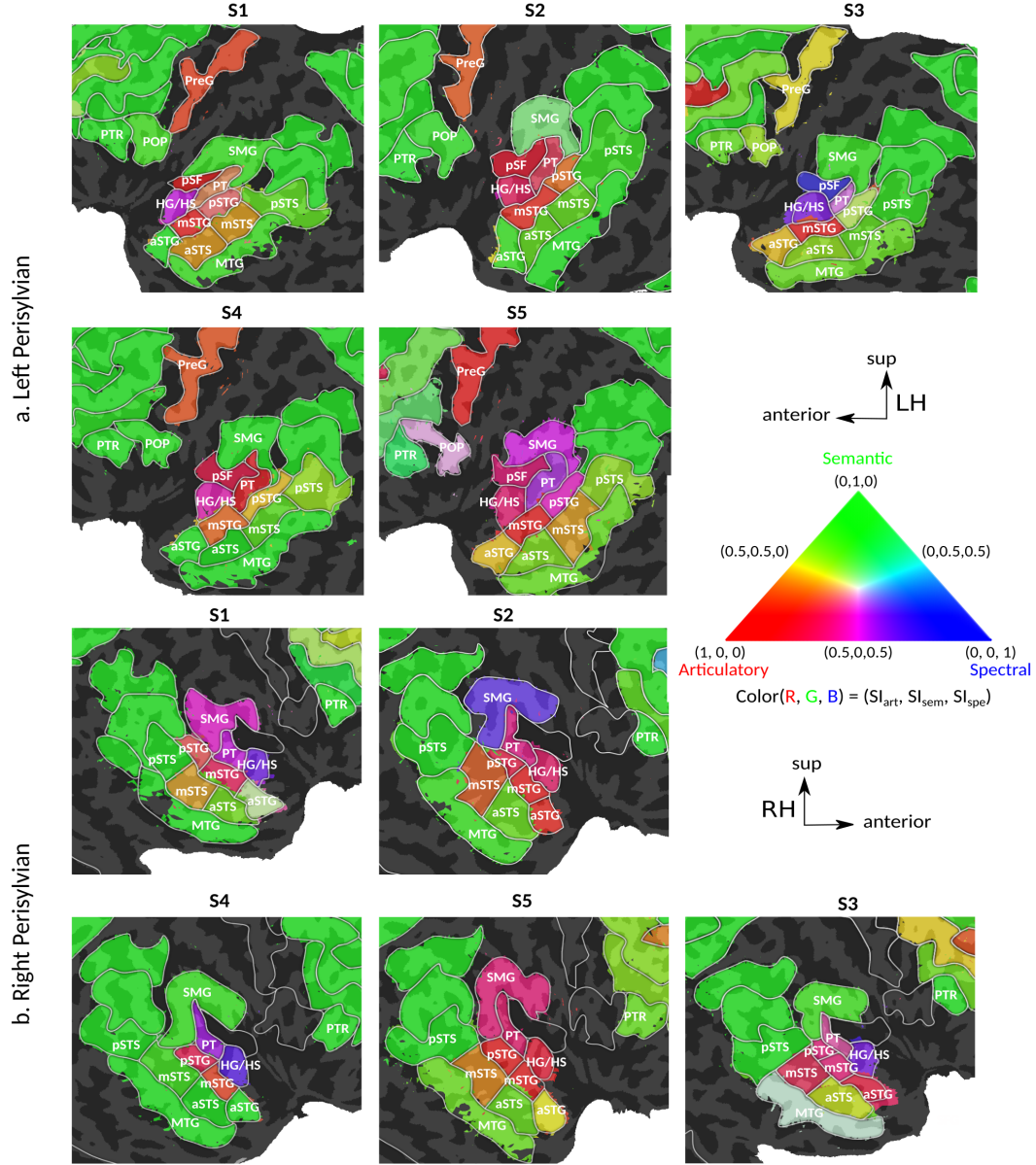


Figure A.6: Intrinsic selectivity profiles. Selectivity profiles of perisylvian ROIs in subjects S1-S5 are shown on the cortical flatmaps. Significant articulatory, semantic, and spectral selectivity indices are projected onto the red, green, and blue channels of the RGB colormap (see Methods). **a.** *Left perisylvian ROIs.* **b.** *Right perisylvian ROIs.* Consistently across subjects, a progression from low- and intermediate-level to high-level speech representations are apparent across bilateral temporal cortex in superior-inferior direction ($HG/HS \rightarrow mSTG \rightarrow mSTS \rightarrow MTG$). These results support the view that speech representations are hierarchically organized across processing hierarchy of speech with partial overlap between spectral, articulatory and semantic representations in early to intermediate stages of auditory processing.

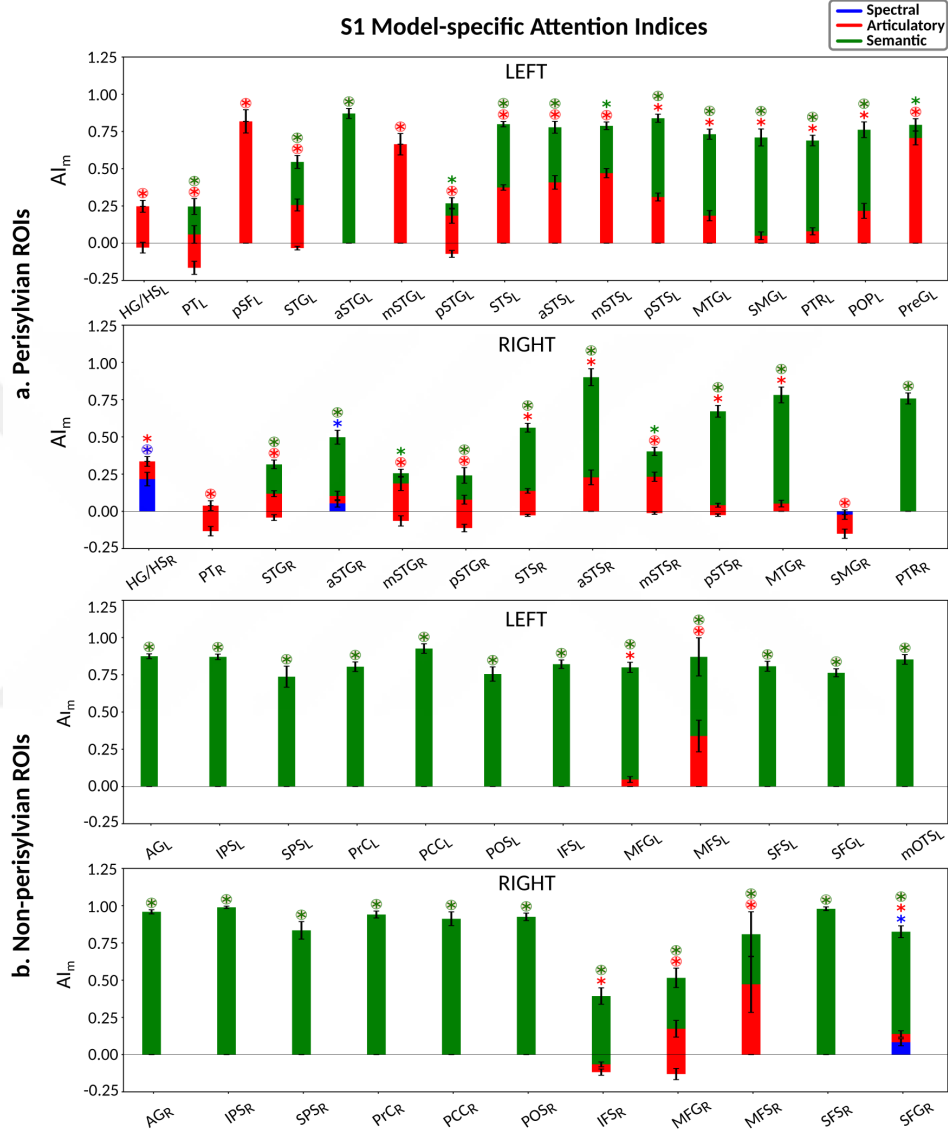


Figure A.7: **Model-specific attention indices for subject S1.** A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across speech-selective voxels in each ROI), and their sum yields the overall modulation. Significantly positive modulations are marked with *, colored according to the corresponding model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a. Perisylvian ROIs.** ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively.

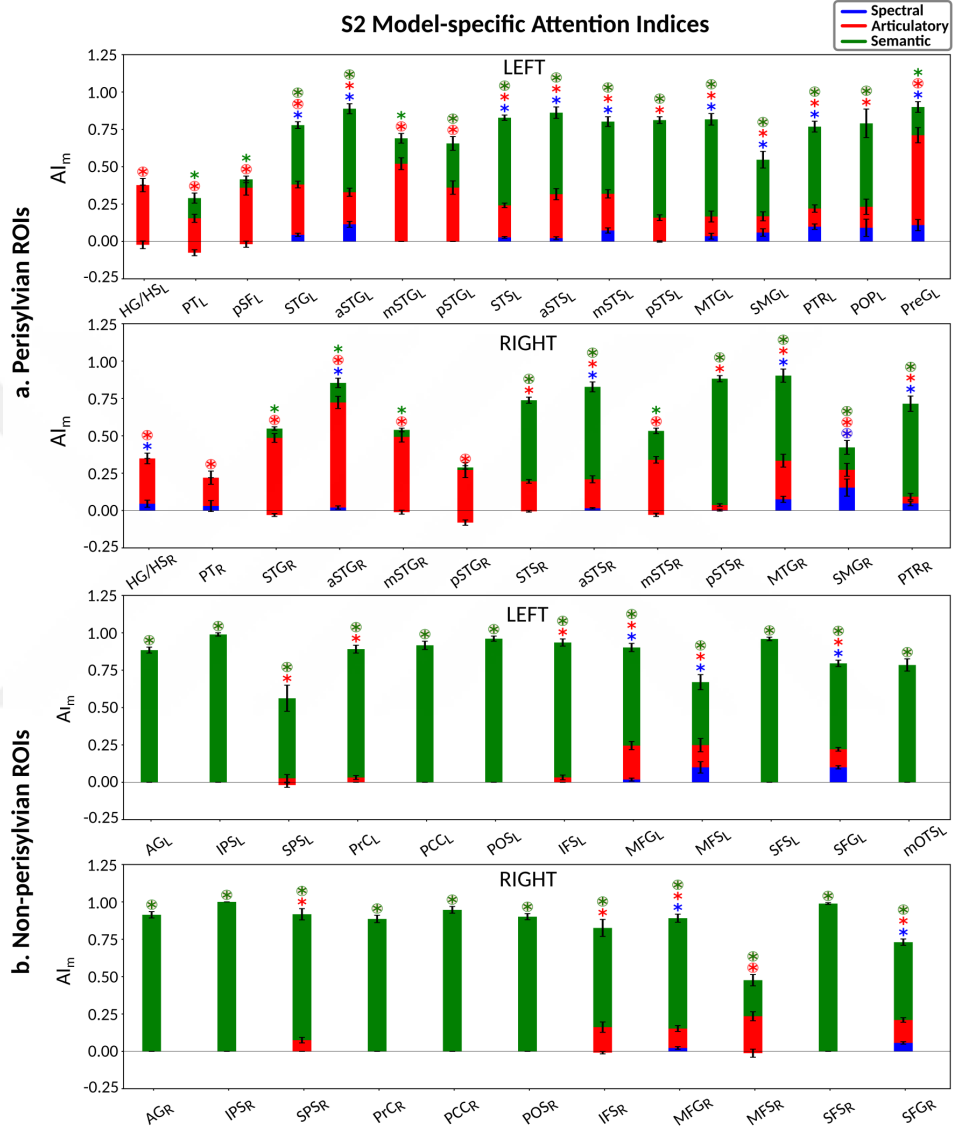


Figure A.8: **Model-specific attention indices for subject S2.** A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across speech-selective voxels in each ROI), and their sum yields the overall modulation. Significantly positive modulations are marked with *, colored according to the corresponding model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a. Perisylvian ROIs.** ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively.

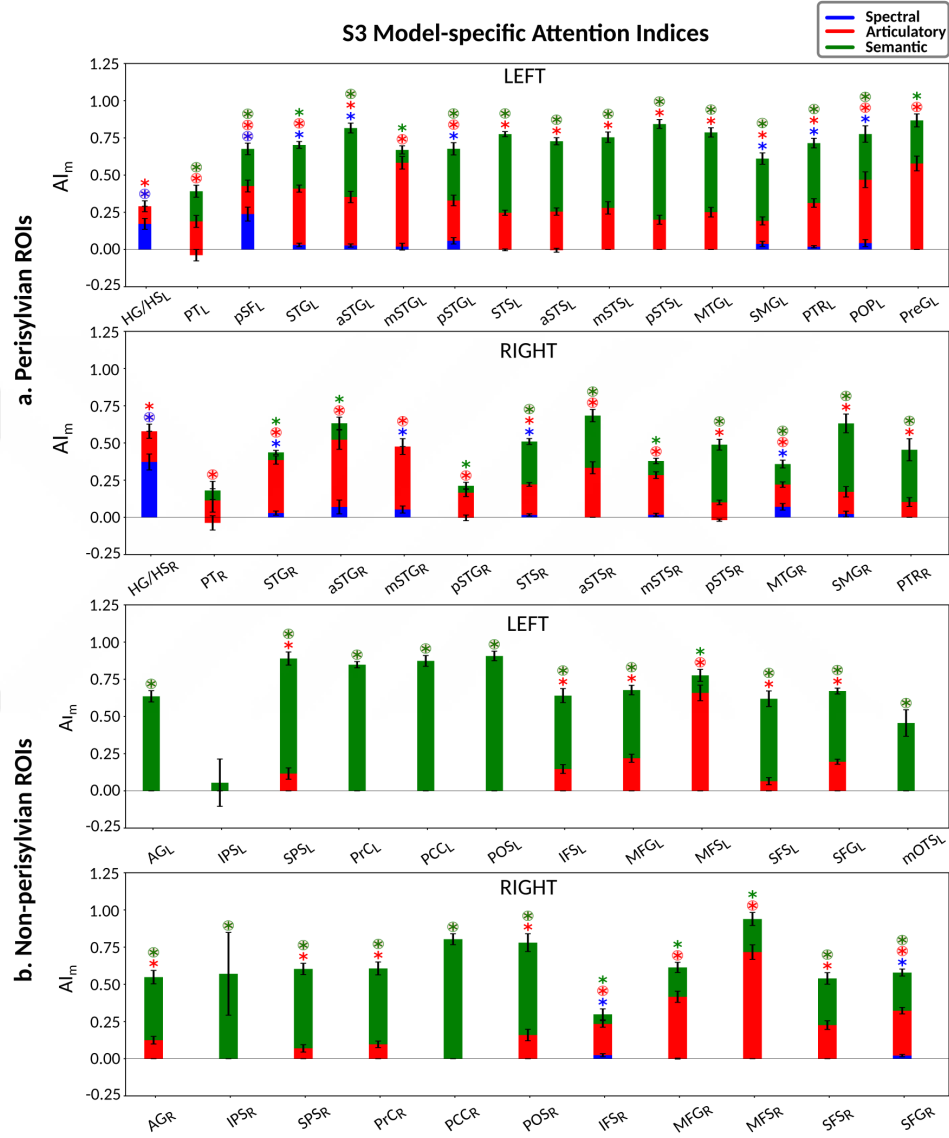


Figure A.9: **Model-specific attention indices for subject S3.** A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across speech-selective voxels in each ROI), and their sum yields the overall modulation. Significantly positive modulations are marked with *, colored according to the corresponding model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a. Perisylvian ROIs.** ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively.

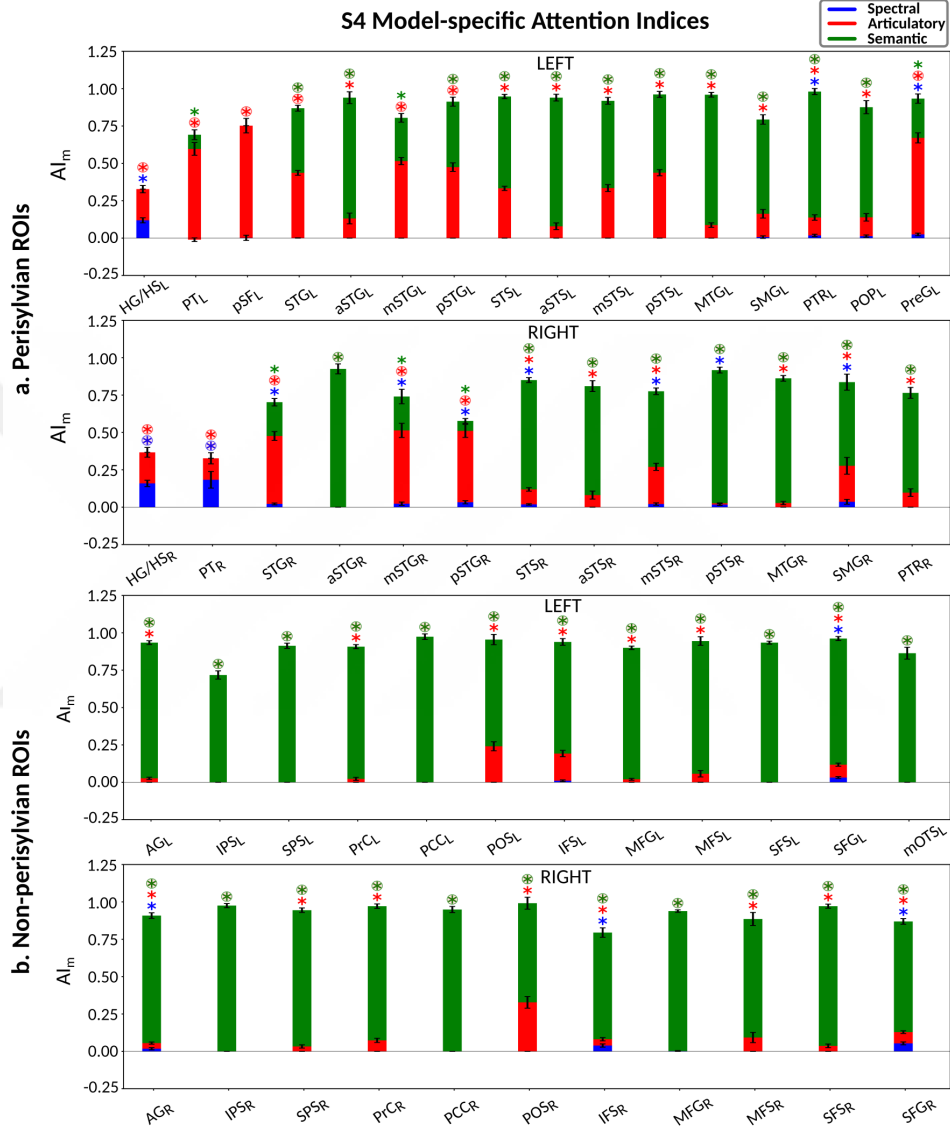


Figure A.10: **Model-specific attention indices for subject S4.** A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across speech-selective voxels in each ROI), and their sum yields the overall modulation. Significantly positive modulations are marked with *, colored according to the corresponding model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a. Perisylvian ROIs.** ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively.

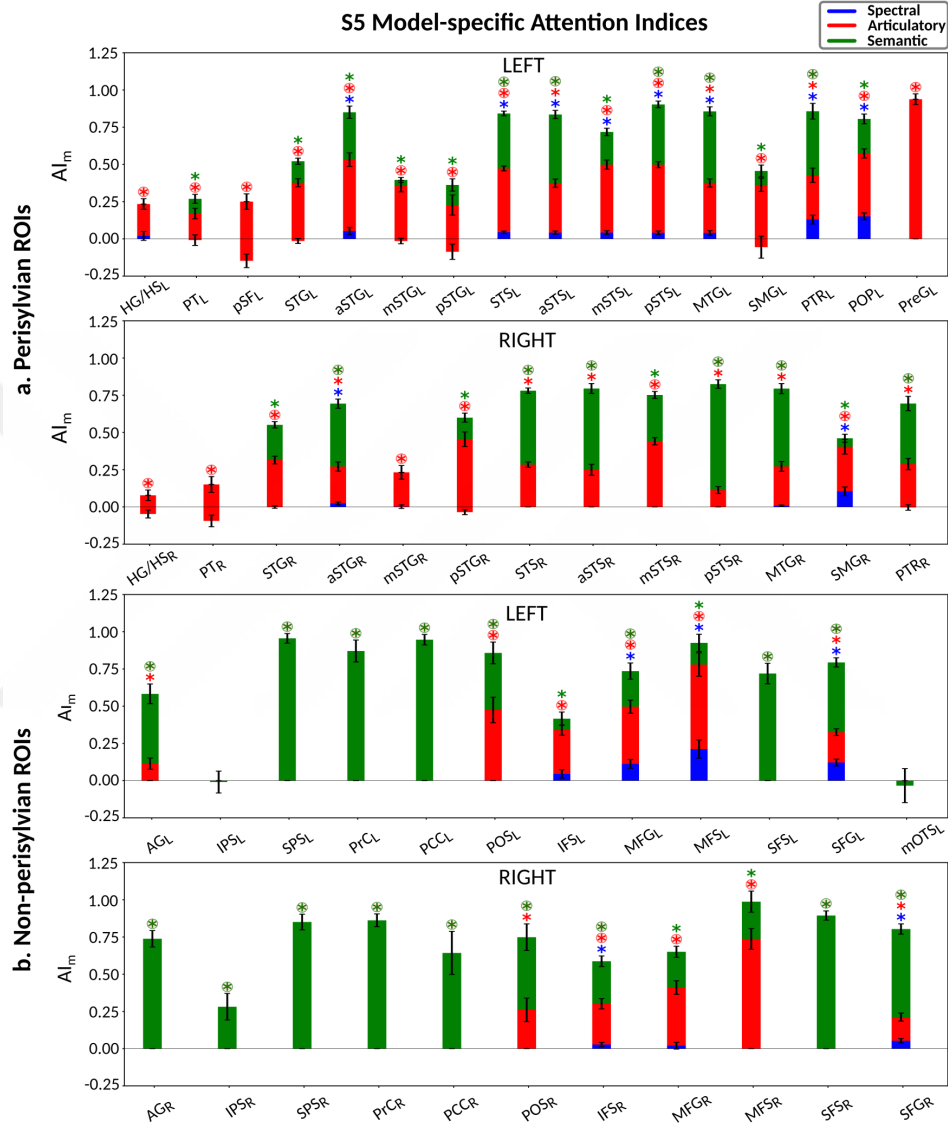


Figure A.11: **Model-specific attention indices for subject S5.** A model-specific attention index (AI_m) was computed based on the difference in model prediction scores when the stories were attended versus unattended (see Methods). AI_m is in the range of $[-1,1]$, where a positive index indicates modulation in favor of the attended stimulus and a negative index indicates modulation in favor of the unattended stimulus. For each ROI, spectral, articulatory, and semantic attention indices are given (mean $\pm$ sem across speech-selective voxels in each ROI), and their sum yields the overall modulation. Significantly positive modulations are marked with *, colored according to the corresponding model and encircled if they are dominant within the ROI ($p < 0.05$, bootstrap test). **a. Perisylvian ROIs.** ROIs in perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively. **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. ROIs in the LH and RH are shown in top and bottom panels, respectively.

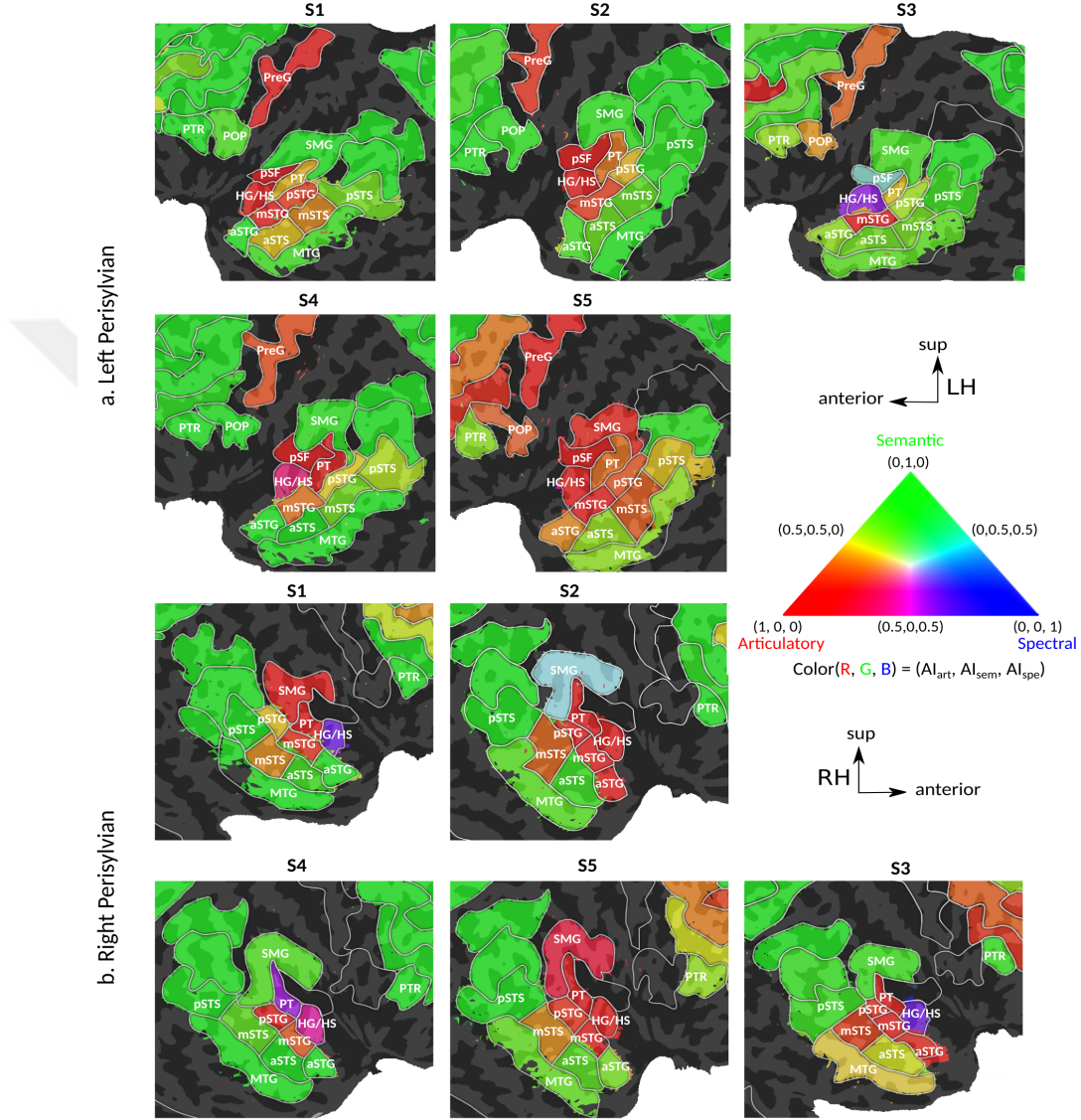


Figure A.12: **Attentional modulation profiles for individual subjects.** Modulation profiles of perisylvian ROIs in subjects S1-S5 are shown on the cortical flatmaps. Significantly positive articulatory, semantic, and spectral attention indices were projected to the red, green, and blue channels of the RGB colormap (see Methods). **a. Left perisylvian ROIs.** **b. Right perisylvian ROIs.** In all subjects, a progression in the level of speech representations dominantly modulated is apparent across bilateral temporal cortex in the superior-inferior direction ($HG/HS \rightarrow mSTG \rightarrow mSTS \rightarrow MTG$). Semantic modulation is dominant at both ends (MTG and PTR) of bilateral ventral stream, while articulatory modulation is dominant at one end (PreG) of left dorsal stream ($p < 0.05$, bootstrap test; see also Supp. Figs. A.7-11). There is no consistent tendency towards dominant articulatory or semantic modulation at the other end of the left dorsal stream (POP).

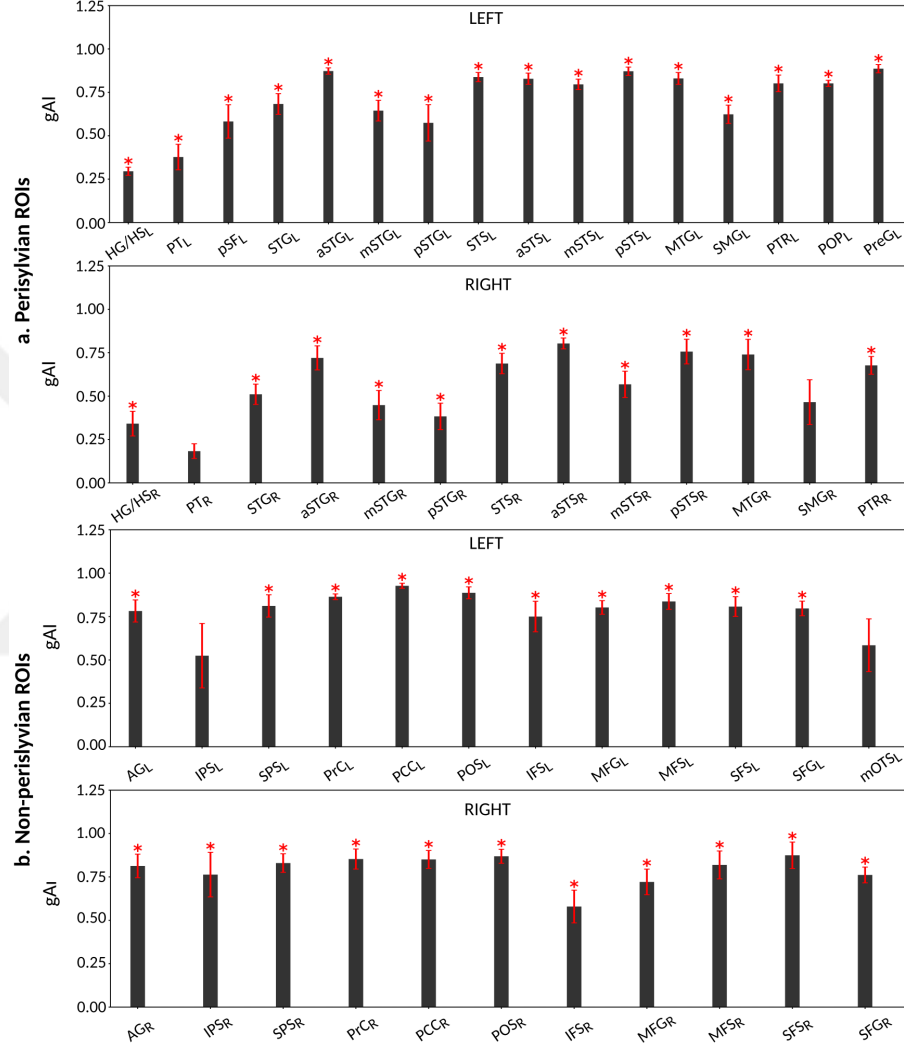


Figure A.13: **Global attention indices.** To quantify overall modulatory effects on selectivity across all examined feature levels, global attentional modulation (gAI) was computed by summing spectral, articulatory, and semantic attention indices (see Methods). gAI is in the range of $[-1,1]$, where a positive index indicates attentional modulation in favor of the attended stimuli and a negative index indicates modulation in favor of the unattended stimuli. A value of zero indicates no modulation. Bar plots indicate gAI (mean $\pm$ sem across subjects). Significant indices are marked with * ($p < 0.05$, bootstrap test; see the bar plots in Supp. Figs. A.7-11 for gAI in individual subjects). **a. Perisylvian ROIs.** ROIs within perisylvian cortex in LH and RH are displayed in top and bottom panels, respectively. gAI is significant in all perisylvian ROIs consistently in all subjects except in right PT and SMG ($p > 0.05$). **b. Non-perisylvian ROIs.** ROIs in non-perisylvian cortex are displayed. gAI is significant in all non-perisylvian ROIs consistently in all subjects except in left IPS and mOTS ($p > 0.05$).

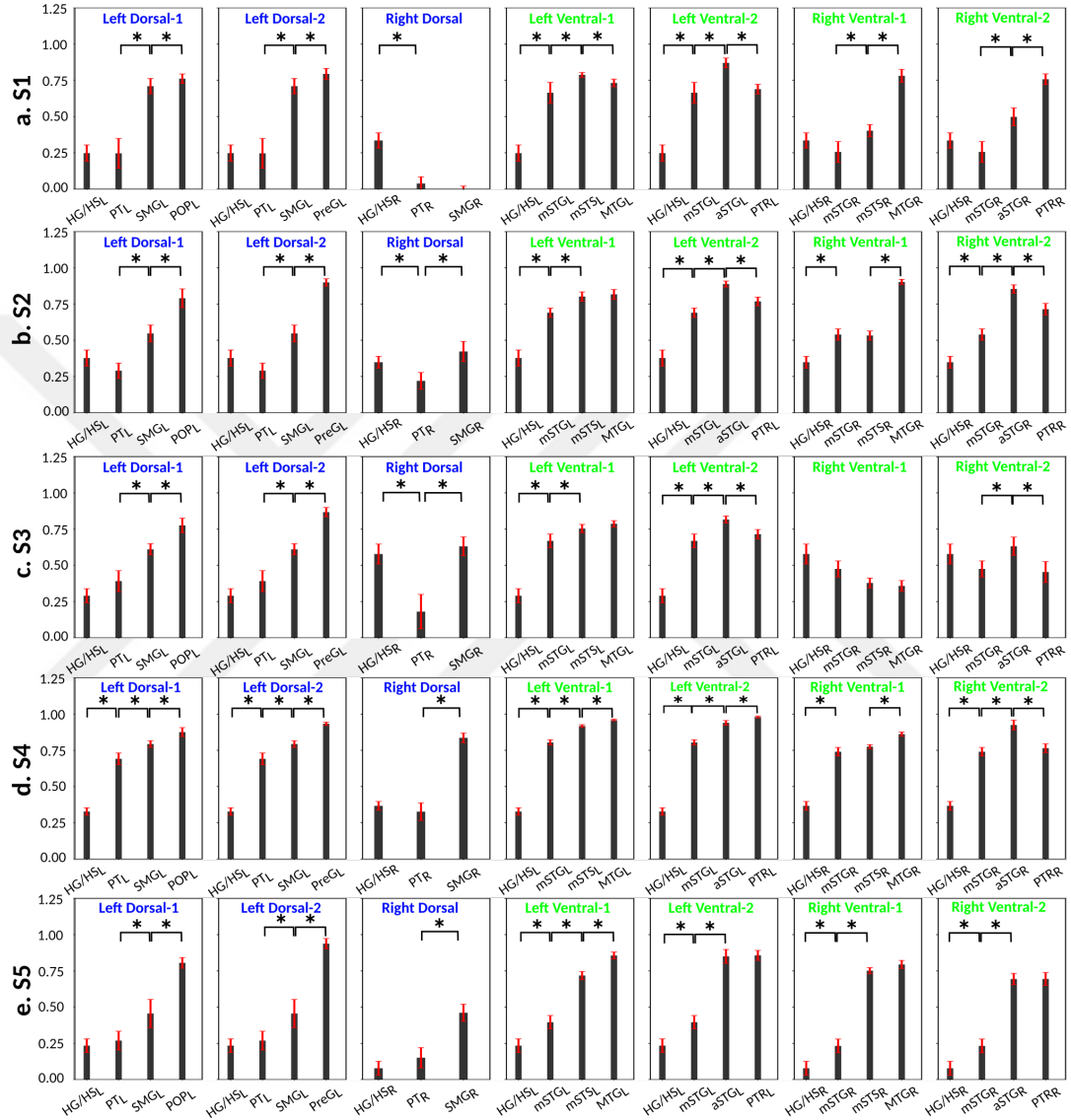


Figure A.14: **Modulation hierarchies for individual subjects.** **a.** *Modulation hierarchies for subject S1.* Gradients in gAI across LD-1, LD-2, RD, LV-1, LV-2, RV-1 and RV-2 are shown in subfigures labeled accordingly. Bar plots display gAI in each ROI (mean $\pm$ sem across speech-selective voxels in each ROI). Only ROIs within a given trajectory are included in the corresponding subfigure. Significant differences in gAI between consecutive ROIs along the trajectory are marked with brackets ($p < 0.05$, bootstrap test). **b.** *Modulation hierarchies for subject S2.* **c.** *Modulation hierarchies for subject S3.* **d.** *Modulation hierarchies for subject S4.* **e.** *Modulation hierarchies for subject S5.* Gradients consistently obtained in each individual subject are ($p < 0.05$): $gAI_{PT} < gAI_{SMG} < gAI_{POP}$ in LD-1, $gAI_{PT} < gAI_{SMG} < gAI_{PreG}$ in LD-2, $gAI_{HG/HS} < gAI_{mSTG} < gAI_{mSTS}$ in LV-1, $gAI_{HG/HS} < gAI_{mSTG} < gAI_{aSTG}$ in LV-2, and $gAI_{mSTG} < gAI_{aSTG}$ in RV-2.

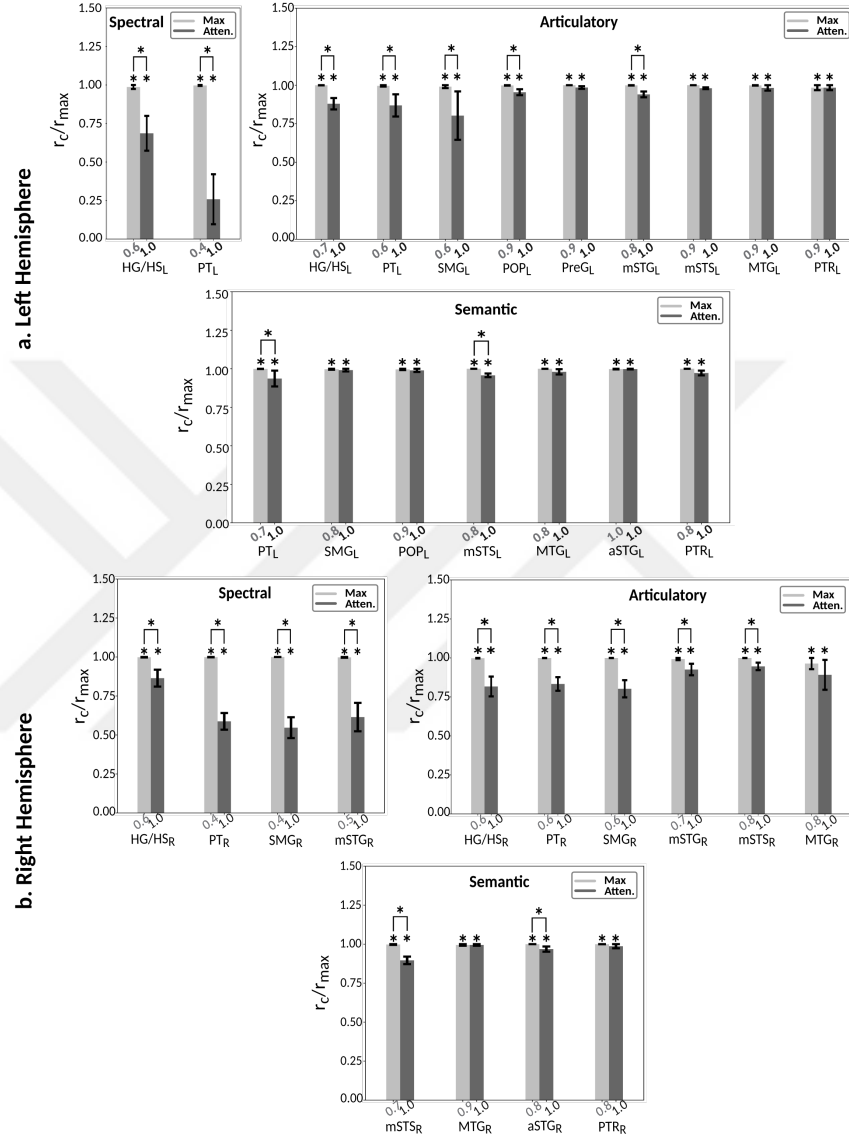


Figure A.15: **Representation of unattended speech in subject S1.** Passive-listening models were tested on cocktail-party data to assess representation of unattended speech during the cocktail-party task. Prediction scores were calculated separately for a combination model comprising features of both attended and unattended stories (r_{max} : optimal convex combination) and an individual model only comprising features of the attended story (r_a). Significant difference in prediction between the two models is an indication that BOLD responses carry significant information on unattended speech. Bar plots display normalized prediction scores (mean $\pm$ sem across speech-selective voxels in each ROI; combination model in light gray and individual model in gray). Significant scores are marked with * ($p < 10^{-4}$, bootstrap test), and significant differences are marked with brackets ($p < 0.05$). Prediction scores are only displayed for ROIs in the dorsal and ventral streams, with significant selectivity for given model features. **a. Left hemisphere.** Spectral, articulatory and semantic representations are labeled in subplots accordingly. **b. Right hemisphere.** Spectral, articulatory and semantic representations are labeled in subplots accordingly.

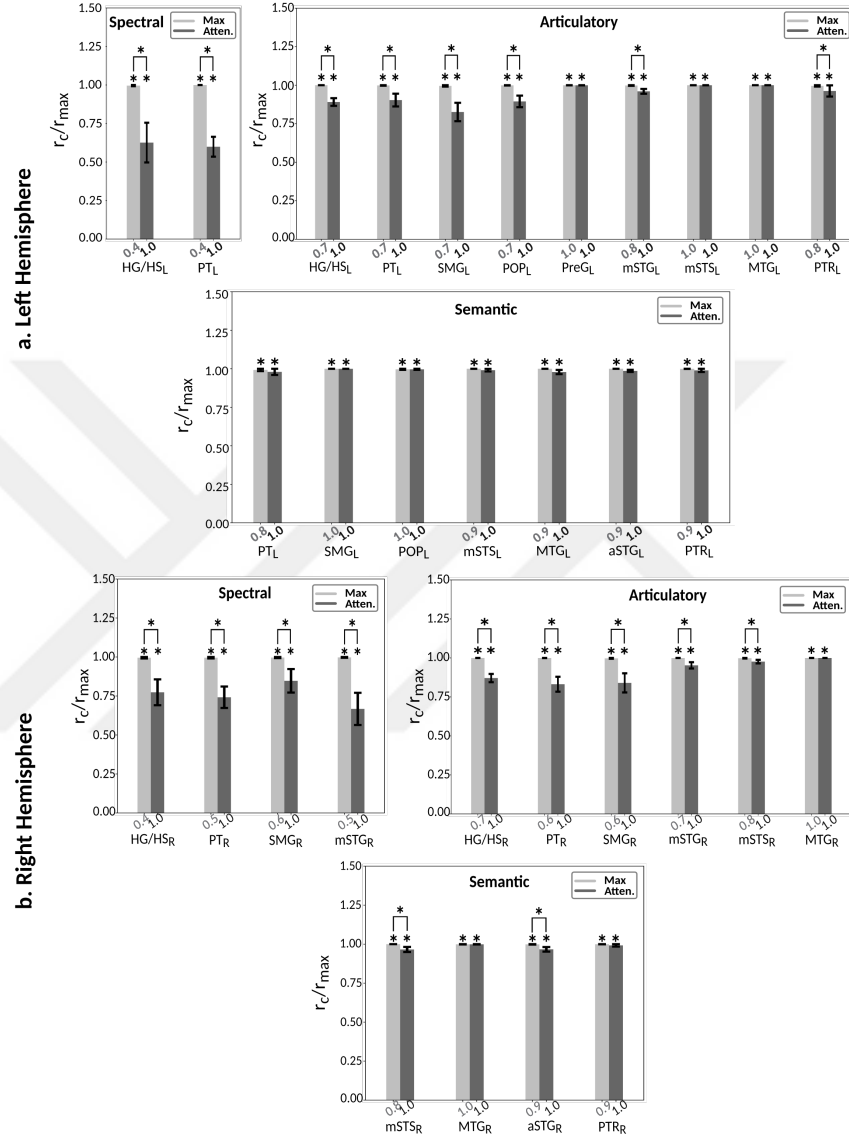


Figure A.16: **Representation of unattended speech in subject S2.** Passive-listening models were tested on cocktail-party data to assess representation of unattended speech during the cocktail-party task. Prediction scores were calculated separately for a combination model comprising features of both attended and unattended stories (r_{max} : optimal convex combination) and an individual model only comprising features of the attended story (r_a). Significant difference in prediction between the two models is an indication that BOLD responses carry significant information on unattended speech. Bar plots display normalized prediction scores (mean $\pm$ sem across speech-selective voxels in each ROI; combination model in light gray and individual model in gray). Significant scores are marked with * ($p < 10^{-4}$, bootstrap test), and significant differences are marked with brackets ($p < 0.05$). Prediction scores are only displayed for ROIs in the dorsal and ventral streams, with significant selectivity for given model features. **a. Left hemisphere.** Spectral, articulatory and semantic representations are labeled in subplots accordingly. **b. Right hemisphere.** Spectral, articulatory and semantic representations are labeled in subplots accordingly.

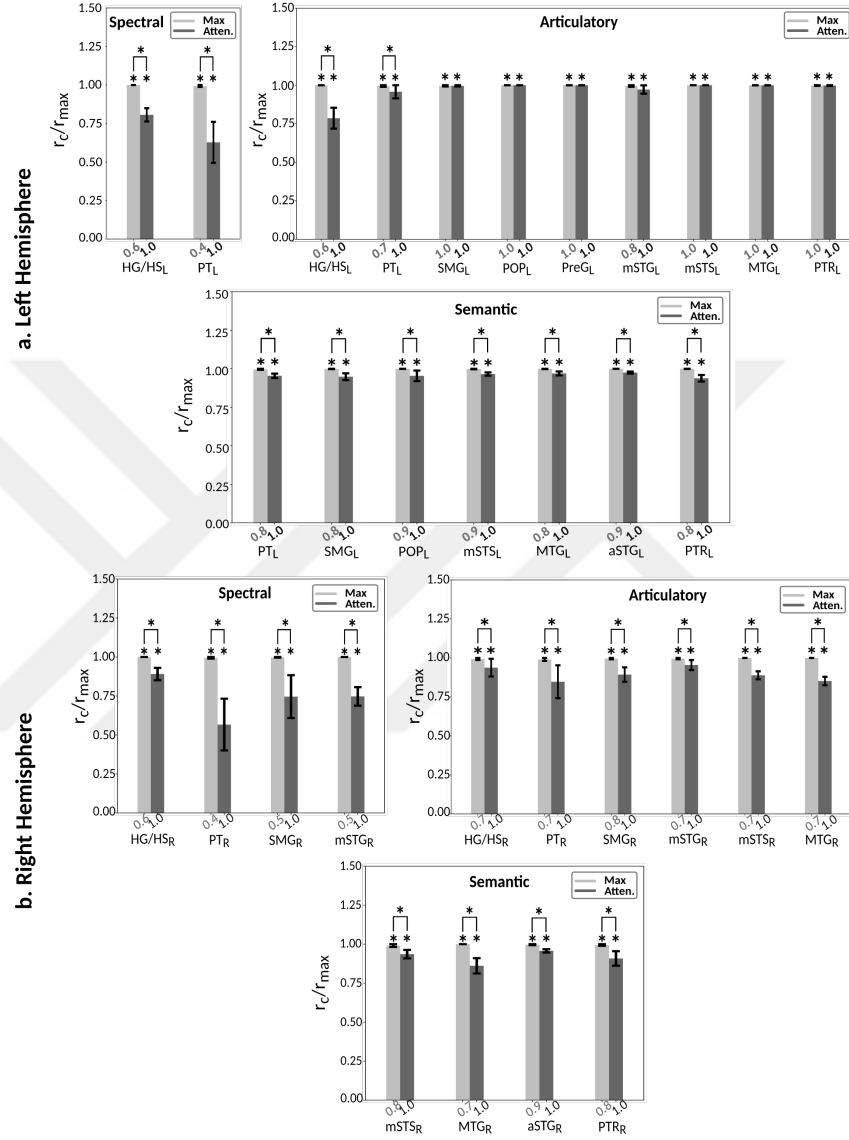


Figure A.17: **Representation of unattended speech in subject S3.** Passive-listening models were tested on cocktail-party data to assess representation of unattended speech during the cocktail-party task. Prediction scores were calculated separately for a combination model comprising features of both attended and unattended stories (r_{max} : optimal convex combination) and an individual model only comprising features of the attended story (r_a). Significant difference in prediction between the two models is an indication that BOLD responses carry significant information on unattended speech. Bar plots display normalized prediction scores (mean $\pm$ sem across speech-selective voxels in each ROI; combination model in light gray and individual model in gray). Significant scores are marked with * ($p < 10^{-4}$, bootstrap test), and significant differences are marked with brackets ($p < 0.05$). Prediction scores are only displayed for ROIs in the dorsal and ventral streams, with significant selectivity for given model features. **a. Left hemisphere.** Spectral, articulatory and semantic representations are labeled in subplots accordingly. **b. Right hemisphere.** Spectral, articulatory and semantic representations are labeled in subplots accordingly.

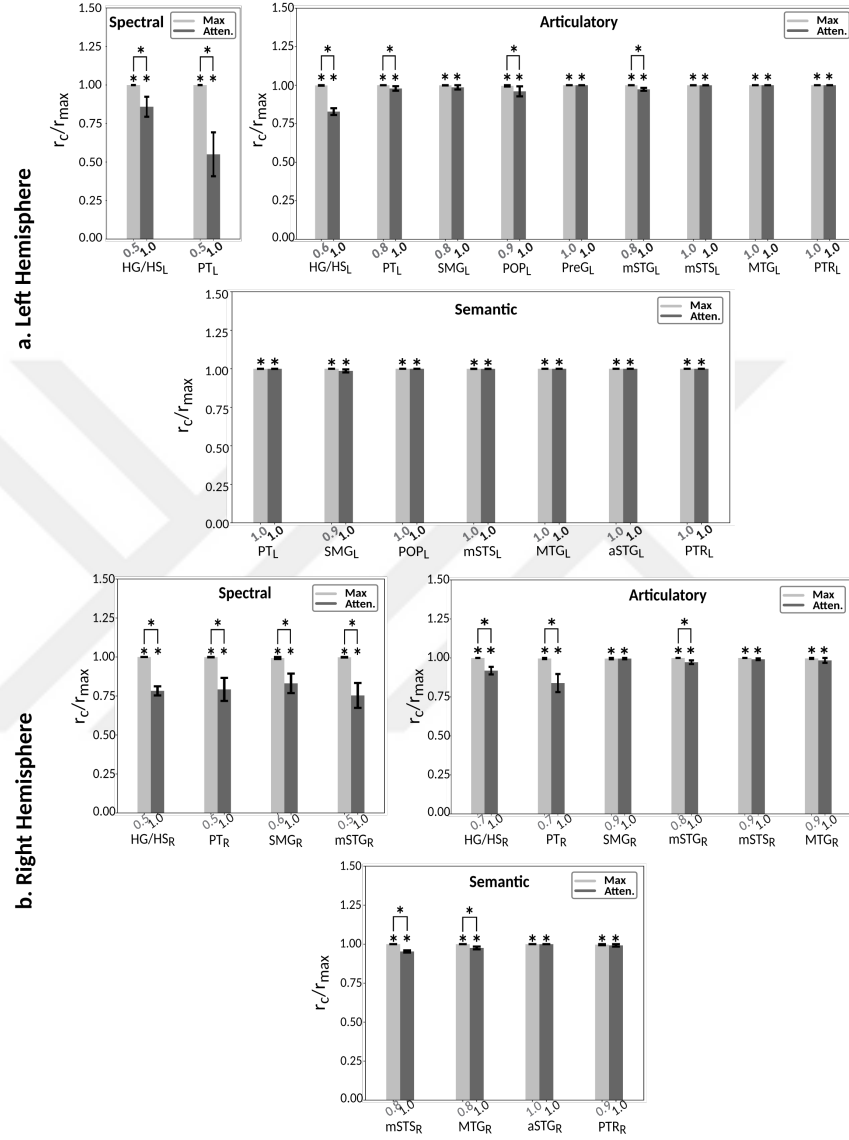


Figure A.18: **Representation of unattended speech in subject S4.** Passive-listening models were tested on cocktail-party data to assess representation of unattended speech during the cocktail-party task. Prediction scores were calculated separately for a combination model comprising features of both attended and unattended stories (r_{max} : optimal convex combination) and an individual model only comprising features of the attended story (r_a). Significant difference in prediction between the two models is an indication that BOLD responses carry significant information on unattended speech. Bar plots display normalized prediction scores (mean $\pm$ sem across speech-selective voxels in each ROI; combination model in light gray and individual model in gray). Significant scores are marked with * ($p < 10^{-4}$, bootstrap test), and significant differences are marked with brackets ($p < 0.05$). Prediction scores are only displayed for ROIs in the dorsal and ventral streams, with significant selectivity for given model features. **a. Left hemisphere.** Spectral, articulatory and semantic representations are labeled in subplots accordingly. **b. Right hemisphere.** Spectral, articulatory and semantic representations are labeled in subplots accordingly.

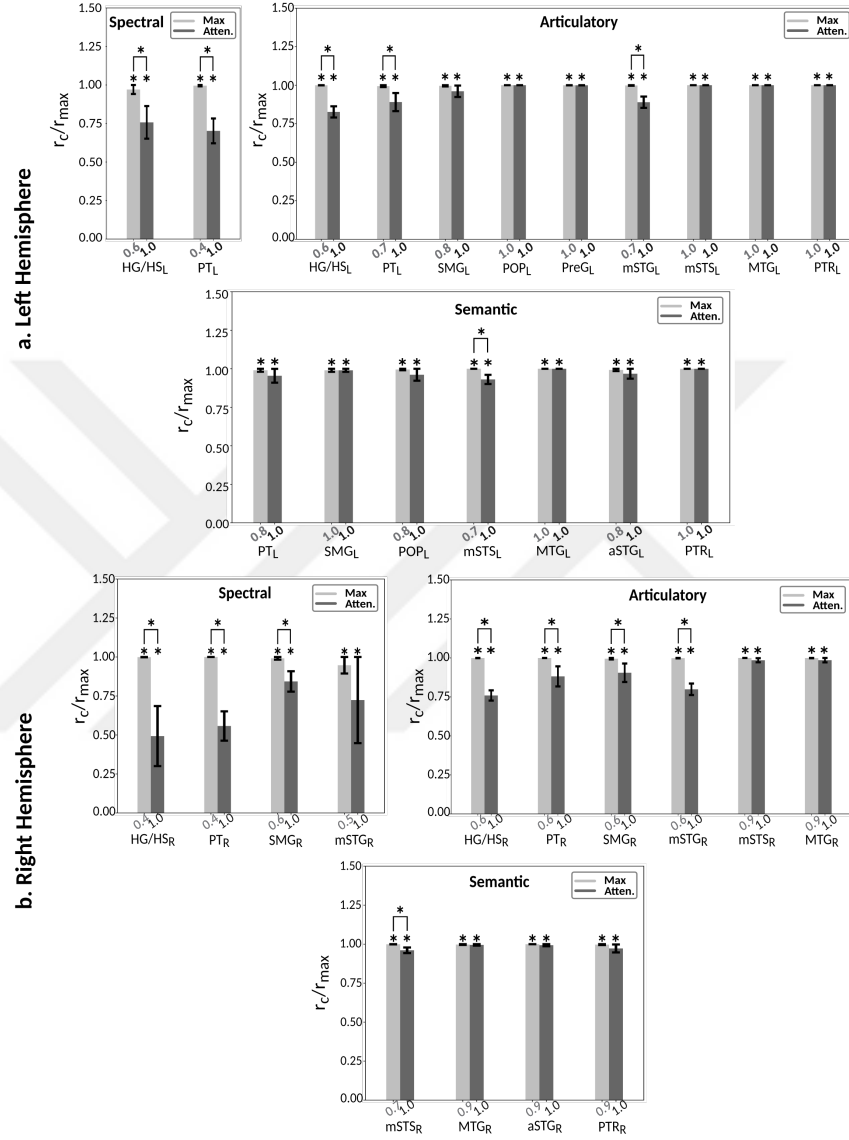


Figure A.19: **Representation of unattended speech in subject S5.** Passive-listening models were tested on cocktail-party data to assess representation of unattended speech during the cocktail-party task. Prediction scores were calculated separately for a combination model comprising features of both attended and unattended stories (r_{max} : optimal convex combination) and an individual model only comprising features of the attended story (r_a). Significant difference in prediction between the two models is an indication that BOLD responses carry significant information on unattended speech. Bar plots display normalized prediction scores (mean $\pm$ sem across speech-selective voxels in each ROI; combination model in light gray and individual model in gray). Significant scores are marked with * ($p < 10^{-4}$, bootstrap test), and significant differences are marked with brackets ($p < 0.05$). Prediction scores are only displayed for ROIs in the dorsal and ventral streams, with significant selectivity for given model features. **a. Left hemisphere.** Spectral, articulatory and semantic representations are labeled in subplots accordingly. **b. Right hemisphere.** Spectral, articulatory and semantic representations are labeled in subplots accordingly.