Published in final edited form as:

Nat Neurosci. 2015 June ; 18(6): 903–911. doi:10.1038/nn.4021.

The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts

Tobias Overath[1,2,6], **Josh H McDermott**[3,6], **Jean Mary Zarate**[2], and **David Poeppel**[2,4,5]

[1]Duke Institute for Brain Sciences, Duke University, Durham, North Carolina, USA.

[2]Department of Psychology, New York University, New York, New York, USA.

[3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, USA.

[4]Center for Neural Science, New York University, New York, New York, USA.

[5]Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany.

Abstract

Speech contains temporal structure that the brain must analyze to enable linguistic processing. To investigate the neural basis of this analysis, we used sound quilts, stimuli constructed by shuffling segments of a natural sound, approximately preserving its properties on short timescales while disrupting them on longer scales. We generated quilts from foreign speech to eliminate language cues and manipulated the extent of natural acoustic structure by varying the segment length. Using functional magnetic resonance imaging, we identified bilateral regions of the superior temporal sulcus (STS) whose responses varied with segment length. This effect was absent in primary auditory cortex and did not occur for quilts made from other natural sounds or acoustically matched synthetic sounds, suggesting tuning to speech-specific spectrotemporal structure. When examined parametrically, the STS response increased with segment length up to ~500 ms. Our results identify a locus of speech analysis in human auditory cortex that is distinct from lexical, semantic or syntactic processes.

The production of spoken language entails encoding thoughts into words and words into sequences of sounds[1]. Listeners must analyze these sounds to derive the speakers' intended meanings. One computational challenge of this task is that the mapping between words and sounds is not one to one: different speakers produce different sounds when expressing the same utterance. As a result, there is reason to think that linguistic analysis cannot proceed

Correspondence should be addressed to T.O. (t.overath@duke.edu) or J.H.M. (jhm@mit.edu).

[6]These authors contributed equally to this work.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS

T.O., J.H.M., J.M.Z. and D.P. designed the experiments, interpreted the data and wrote the manuscript. J.H.M. designed the quilting algorithm and generated the stimuli. T.O. and J.M.Z. acquired the fMRI data. J.H.M. acquired the behavioral data. T.O. and J.H.M. analyzed the data.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

directly from the sound waveform itself, but rather might be preceded by a stage of acoustic analysis that maps patterns of sound energy onto intermediate, invariant representations of features, phonemes or syllables[2]. Here we identify a locus of speech-specific acoustic analysis that is a candidate precursor to linguistic processing.

Although studies of speech perception have generally highlighted neural substrates in the superior temporal lobes[3–8], much about the auditory analysis of speech remains poorly understood. In particular, prior work has generally not addressed the existence and nature of auditory mechanisms for speech analysis *per se* (that is, for analyzing acoustic rather than linguistic structure). One influential approach has been to manipulate speech in ways that affect its intelligibility, typically changing both acoustic and linguistic content[3,9–11] and leaving open the question of whether acoustic and linguistic processing are distinct in cortex. Other studies have targeted speech-relevant acoustic processing using synthetic non-speech stimuli[12–16], leaving open whether the implicated regions and mechanisms are speech specific.

Our goal was to isolate the auditory analysis of speech, particularly its temporal attributes, by manipulating foreign speech that had no lexical-semantic or syntactic content for our listeners. The premise of our approach is that one might expect neurons underpinning speech analysis to be tuned to the particular structures that occur in natural speech, such that they would respond more to sound signals that have naturalistic speech structure than to those in which this naturalistic structure is disrupted. Notably, speech is richly structured in time, with systematic organization at multiple timescales[1,17,18]. Vowels and consonants can be distinguished by spectral changes over tens of milliseconds, syllables by sequences of phonemes over a few hundred milliseconds, and the intonational contours of phrases and sentences by pitch variation over hundreds to thousands of milliseconds. We manipulated acoustic structure on particular timescales by dividing a natural sound into segments and reordering them in a manner subject only to local constraints (Fig. 1a and Online Methods). We call such stimuli ‘sound quilts’. The synthesis methodology was inspired by methods in image processing that synthesize images by ‘stitching together’ patches of a source image[19]. Quilt segments are ordered and appended such that segment-to-segment changes resemble those in the source signal, but are otherwise unconstrained. The structure of a quilted signal is therefore similar to that of the source signal within a segment and at a segment’s border, but differs from the source at larger scales for source signals that contain large-scale dependencies.

We hypothesized that regions subserving any putative speech-specific analysis would exhibit an increasing response to speech quilts as the segment length was increased, as this manipulation increases the temporal extent over which the signal contains naturalistic speech structure. It also seemed plausible that responses should be limited by the analysis timescale of the part of the auditory system under consideration: if two different segment lengths both exceed the temporal integration window of a region’s characteristic neuronal receptive fields, the response to quilts composed of the two segment lengths should be similar, as they will appear to be similarly natural from the perspective of the temporal receptive field. Responses to quilts with different segment lengths were therefore intended to both identify the location of potentially speech-specific analysis mechanisms and to help

characterize the timescale of their analysis. To distinguish the analysis of speech acoustics from responses that might be driven by lexical, syntactic, or semantic structure, we generated quilts from speech in a language (German) that was foreign to our English-speaking participants but that has considerable phonological overlap with English, and would therefore engage speech analysis mechanisms that normally process English.

Our approach is conceptually similar to a study that used image scrambling to probe neuronal tuning for structure at different spatial scales in the visual system[20], as well as to studies that have used scrambling stories to examine the timescale of narrative representation[21] and variable length word lists to examine linguistic processes[22]. Other studies have scrambled music and speech on one particular (long) timescale (for example, see ref. 23). However, randomly re-ordering sound or image segments typically introduces perceptually salient boundary artifacts; our quilting algorithm minimized these by ordering segments to match the degree of segment-to-segment change in the original signal and by concatenating them to avoid signal discontinuities.

RESULTS

We analyzed the BOLD signal in regions of interest (ROIs) defined either functionally or anatomically. As a functional localizer, we contrasted the response to quilts composed of long segments with the response to quilts composed of short segments (960 and 30 ms, respectively, the longest and shortest segments used; Fig. 1b). This contrast identifies voxels that respond significantly more to signals with naturalistic structure (at a range of timescales up to approximately 1 s) than signals that lack such structure. Activations for this contrast were located bilaterally in the STS (Fig. 2a) and were evident in every individual participant that we tested (Fig. 2b). We also assessed the reverse contrast, as described below, but this did not yield significant activations.

We used this localizer contrast to define a functional ROI (fROI) that was sensitive to quilt segment length and then measured the response to additional conditions in this independently defined fROI. In our initial analyses, we examined fROIs derived from the localizer contrast, applied either in individual participants (individual fROI) or across the entire group of participants (group fROI). The group fROI, resulting from a group second-level random effects analysis of the functional localizer contrast, was included in part for comparison with the anatomical ROIs, which were generated from standard probabilistic maps of anatomical variation across large groups of participants.

To investigate the timescales over which these regions were sensitive to acoustic structure, we parametrically varied the quilt segment length from 30 to 960 ms and measured the response in our ROIs. To facilitate comparisons across different ROIs (whose overall response often varied considerably), we normalized the response to each condition in an ROI by its response to the 960-ms localizer condition. The average response to speech quilts in both individual and group fROIs increased with segment length up until 480 ms, at which point it plateaued (main effect of segment length in both cases; individual, $F_{(3.01,42.18)} = 112.42$, $P < 0.001$, $\eta^2_p = 0.89$; group, $F_{(5,70)} = 28.0$, $P < 0.001$, $\eta^2_p = 0.67$; Fig. 3).

The effect size was slightly larger in the left hemisphere (the main effect of hemisphere was marginally not significant: $F_{(1,14)} = 4.63$, $P = 0.05$), producing significant interactions between the effects of segment length and hemisphere (individual, $F_{(1.99,27.82)} = 13.43$, $P < 0.001$, $\eta^2_p = 0.49$; group, $F_{(2.85,39.95)} = 4.04$, $P < 0.05$, $\eta^2_p = 0.22$). Similarly, the extent of the fROIs in individual subjects showed a trend toward left-lateralization: a paired samples t test on the volume of the individual fROIs between the two hemispheres yielded a nearly significant difference ($t_{(15)} = 2.03$, $P = 0.06$). Overall, however, the effect was robust in both hemispheres.

In contrast with the fROIs, responses in Heschl's Gyrus (HG, inclusive of primary auditory cortex) and planum temporale (PT, part of non-primary auditory cortex) showed substantially less variation with segment length (as expected, given that these regions were largely outside the fROIs; Fig. 2a). HG exhibited a weak main effect of quilt segment length ($F_{(5,70)} = 2.5$, $P = 0.04$, $\eta^2_p = 0.15$), but PT did not, and no pairwise comparisons (Bonferroni corrected) between any pair of segment lengths were significant in either hemisphere of either ROI ($P > 0.05$ in all cases). An ANOVA across all ROIs accordingly showed an interaction between the effect of quilt segment length and ROI ($F_{(15,210)} = 49.17$, $P < 0.001$, $\eta^2_p = 0.78$).

We also tested for areas that responded more to short-segment speech quilts than long-segment speech quilts via the reverse functional localizer contrast [L30 > L960]. Most participants did not exhibit any significant activations in the superior temporal gyrus (STG) for this contrast, and no voxels in auditory cortex were significantly activated at the group level (all $P > 0.001$, uncorrected; Supplementary Fig. 1).

These results suggest some specialization for the processing of acoustic structure in speech: regions in STS showed strong sensitivity to the temporal extent of natural speech structure that was not mirrored in regions that lie earlier in the presumptive cortical processing hierarchy. Because there were no obvious qualitative differences between the individual and group fROIs, we restricted subsequent analyses to the individual fROIs, which best identified the locus of our key effect, and to the two anatomical ROIs HG and PT, which reveal the response in putatively earlier primary and non-primary auditory cortex, respectively.

Tests of speech specificity

Do the responses to speech quilts reflect speech-specific processing? To rule out more generic explanations of our results, we measured responses to quilts composed of various control stimuli. In each case, we presented control quilts made from long (960 ms) and short (30 ms) segments, included as additional conditions in sessions in which we also measured the parametric response to speech quilts. For ease of comparison, the responses to the control conditions are plotted with responses to speech quilts of the same segment lengths (using data independent of that used to define the fROIs).

Modulation control stimuli

We first sought to test whether the results could be explained by responses to basic amplitude modulation characteristics. Quilting produces stimuli whose long-term power spectra are similar to that of the source signal, irrespective of the segment length (Supplementary Fig. 2), but the modulation spectra exhibit some variation with segment length. For speech, quilts with short segments yielded somewhat less power at rates associated with syllable alternation (3–5 Hz) than did quilts with longer segments (Fig. 4a). In principle, differences in modulation spectra could at least partially account for the lower response to short-segment quilts[14–16,24,25]. To control for these differences in modulation, we synthesized stimuli that replicated them. Specifically, we used the McDermott and Simoncelli sound texture synthesis algorithm to decompose speech signals using an auditory model, measure the marginal moments of the envelopes of each frequency channel as well as the power in a set of modulation filters, and then synthesize stimuli with the same marginal moments and modulation power distributions[26]. We then generated quilts from these synthetic signals in the same way that we did for speech. These modulation-control quilts have modulation spectra (and power spectra; Supplementary Fig. 2) that are similar to those of speech quilts, although they are otherwise unconstrained and do not resemble speech (Figs. 1b and 4a).

In the fROI, responses to the modulation-control stimuli were weak overall, far below even the response to 30-ms speech quilts, and were similar for short and long segments (Fig. 4b; main effect of quilt type, $F_{(1,8)} = 366.71$, $P < 0.001$, $\eta^2_p = 0.98$; with an interaction between quilt type and segment length, $F_{(1,8)} = 26.6$, $P = 0.001$, $\eta^2_p = 0.77$; the effect of segment length in the control quilts alone was not significant, $P > 0.05$). In contrast, responses to the control stimuli in HG and PT were comparable to those for the speech quilts (there were no main effects or interactions for quilt type and segment length in HG or PT, all $P > 0.05$), showing that the larger response to speech quilts in the fROI was not present throughout the auditory cortex. These results indicate that the response to the speech quilts was not being driven by their within-channel modulation spectra, as these were matched in the control stimuli.

As a further test, we examined whether the responses to speech quilts might be driven by correlated modulation across frequency channels (co-modulation), a common property of natural sounds that our modulation control stimuli lacked (Fig. 4c). We generated an additional set of control stimuli in which cross-channel correlations (as well as envelope marginal moments and modulation power distributions) were matched to those of speech, using the same statistical synthesis procedure[26]. Quilts made from these stimuli again showed a low overall response in the fROIs and, at best, a weak effect of segment length (Fig. 4d), producing a main effect of quilt type ($F_{(1,4)} = 98.09$, $P = 0.001$, $\eta^2_p = 0.96$) and an interaction between quilt type and segment length ($F_{(1,4)} = 78.83$, $P = 0.001$, $\eta^2_p = 0.95$) (the effect of segment length was not significant for the control quilts by themselves, $P > 0.05$). The response to these control quilts in the anatomical ROIs of HG and PT was again comparable to that for speech quilts. There was no effect of quilt type or segment length in HG (all $P > 0.05$), and although the response in PT showed a main effect of quilt type ($F_{(1,4)} = 14.77$, $P < 0.05$, $\eta^2_p = 0.79$; all others $P > 0.05$), this was driven by a slightly higher

response to the control stimuli. Taken together, these two control experiments indicate that generic responses to amplitude modulation cannot account for the responses to speech quilts.

Environmental sounds

Are the brain regions that are sensitive to the acoustic structure of speech also sensitive to the structure of other natural sounds? To further address the speech-specificity of our effects, we examined responses to quilts made from a set of environmental sounds. The sounds encompassed animal vocalizations (dogs barking, bird songs) and human actions (footsteps, sawing wood) selected to have structure over long timescales (Online Methods). We generated quilts using the same procedure used for speech.

Quilts made from environmental sounds evoked an overall response in the individual fROIs that was much lower than that for any of the speech quilts, irrespective of segment length (main effect of quilt type: $F_{(1,4)} = 72.59$, $P = 0.001$, $\eta^2_p = 0.95$; Fig. 5). The effect of segment length on the fROI response was also much larger for speech quilts than for environmental sound quilts, producing an interaction between quilt type and segment length ($F_{(1,4)} = 9.45$, $P < 0.05$, $\eta^2_p = 0.7$), although a pairwise comparison between the two environmental sound quilt control conditions was significant ($P < 0.05$). In contrast, the response in HG showed a weak main effect of segment length ($F_{(1,4)} = 15.59$, $P < 0.05$, $\eta^2_p = 0.8$), but no main effect of quilt type or interaction between quilt type and segment length. The response in PT showed a weak main effect of quilt type ($F_{(1,4)} = 8.9$, $P < 0.05$, $\eta^2_p = 0.69$) and a main effect of segment length ($F_{(1,4)} = 36.24$, $P < 0.01$, $\eta^2_p = 0.9$), but no interaction, and it is apparent that the main effect of quilt type is driven by a larger response to the environmental sound quilts. This effect contrasts with the much lower response to the same quilts observed in the fROI. These results further support the claim that there is functional specialization for speech: the overall fROI response to quilts and the effect of segment length were substantially weaker for environmental sounds than for speech, and both of these differences were not present in other auditory areas.

Noise-vocoded speech

What aspects of speech drive the response to speech quilts? One possible account is that the responses are driven by prosody, particularly the prosodic pitch variation that occurs in natural speech. Such patterns are largely preserved in quilts made from long segments of speech, but are severely disrupted in quilts made from short segments. To test the importance of pitch variation, we presented quilts made from noise-vocoded speech[27]. We generated noise-vocoded versions of each of our source speech recordings by imposing the envelopes of ten frequency bands (constructed to be equally spaced along the cochlea, covering the audible spectrum) on noise (Fig. 1b). Pitch, which relies on fine spectral detail, is eliminated by this procedure, but coarse spectral content sufficient for phonetic identification remains present. When English speech is noise-vocoded in this way, intelligibility remains high[27,28].

The fROI response to quilts of noise-vocoded speech was somewhat lower than that to normal speech quilts, but showed a comparable effect of segment length (main effect of segment size, $F_{(1,4)} = 21.97$, $P < 0.01$, $\eta^2_p = 0.85$, without an interaction with quilt type; Fig.

6). Although the response to normal speech quilts was stronger than that to noise-vocoded speech, producing a main effect of quilt type ($F_{(1,4)} = 91.55$, $P = 0.001$, $\eta^2_p = 0.96$), responses to the noise-vocoded stimuli were nonetheless much higher than those to the various non-speech quilts that we tested (Figs. 4 and 5). Separate repeated-measures ANOVAs for the response in HG and PT, in contrast, revealed no main effects or interactions (all $P > 0.05$). These results suggest that the effect of segment size for quilts of normal speech is not driven by pitch variation. Rather, other aspects of speech-specific temporal structure are evidently important.

Timescale of acoustic analysis

One notable feature of the parametric response to speech quilts (Fig. 3) is the response plateau at 480 ms. Although individual participants exhibited some variability in their response profile, an elbow at 480 ms was visually evident in most cases (Supplementary Fig. 3). To quantitatively evaluate the response shape, we fit the parametric responses with a piecewise linear function consisting of an initial portion of variable slope and a later portion that was flat. The functions were parameterized by the response at 30 ms, the slope of the initial portion and the position of the elbow connecting the sloping and flat portions. The functions could therefore accommodate elbows in different positions as well as different overall levels of response. To maximize our power to resolve the location of an elbow point, we averaged responses across hemispheres.

To evaluate whether the elbow function better described the parametric response than did a linear function, we measured the average error of both functions on left-out data. For each participant, we fit both function types to the parametric fROI response averaged across the other 14 participants and then measured the residual error in the left-out participant's data. Such a procedure automatically penalizes model complexity, as complexity that is not needed to account for structure in the data produces over-fitting. Instead, we found that the elbow model produced consistently lower error (mean r.m.s. error on left-out data of 0.065 for the elbow model versus 0.069 for the linear model, significantly different via a paired t test: $t(14) = 3.17$, $P = 0.0068$).

One potential explanation for this response plateau is that the regions that we identified integrate sound information at timescales up to about half a second. They respond more to quilts composed of 480-ms segments than quilts composed of 240-ms segments (and to 240 versus 120, 120 versus 60, and 60 versus 30 ms), suggesting that they are sensitive to differences in naturalistic acoustic structure between these timescales. In contrast, quilts with 960 ms and 480 ms segments produce largely equivalent responses. This result is what one might expect if temporal receptive fields in these regions extended up to about 500 ms in duration.

An alternative explanation is that the response plateau reflects the statistics of speech. Specifically, it is conceivable that (foreign) speech does not contain substantial acoustic dependencies past half a second, such that 480 ms and 960 ms quilts are largely acoustically equivalent. As a first step toward addressing this issue, we presented listeners with speech quilts of different segment lengths and asked them to rate the extent to which they sounded natural. These subjective ratings showed a pattern that was distinct from that of the fROI

response: rated naturalness increased monotonically with segment length, without an elbow at 480 ms ($F_{(2.73,40.8)} = 455.54$, $P < 0.001$, $\eta^2_p = 0.97$, with significant pair-wise comparisons between all adjacent conditions; Fig. 7a). These results indicate that there is discriminable acoustic structure in (foreign) speech at timescales beyond 480 ms. This structure is apparent to listeners, but does not alter the overall neural response magnitude in the fROI.

Compressed speech

To further explore the origins of the parametric response, we measured responses to quilts created from time-compressed speech. We used the original German speech recordings and compressed them by a factor of two. We reasoned that such signals contain twice as much speech information in a segment of a given length as uncompressed (normal) speech. If the parametric response in Figure 3 merely reflects dependencies in the stimulus, the response to quilts of compressed speech should shift leftward. Specifically, if the response plateau is a result of the dependencies in normal speech being negligible beyond 480 ms, such that quilts with 480 ms segments are acoustically equivalent to quilts with longer segments, the same phenomenon should occur at 240 ms in compressed speech, as the temporal dependencies that are present in 480 ms of normal speech are packed into 240 ms of compressed speech. If, instead, the response plateau reflects an intrinsic timescale of analysis in the auditory system (receptive field time constants, for instance), then the dependence of the response on segment length might be expected to remain the same. Under this latter hypothesis, assuming that the brain is sensitive to the structure in compressed speech despite the time-compression, the fROI response should increase with segment length up to the temporal limits of its analysis. To minimize other differences between stimuli, we performed compression with an algorithm that preserved pitch, such that the fundamental frequency range was the same for regular and time-compressed speech stimuli.

Although the fROI response to quilts of compressed speech was weaker than to those of normal speech, the shape of the parametric response was similar (Fig. 7b). The response in HG, in contrast, was similar for quilts of compressed and uncompressed speech, and a repeated-measures ANOVA with factors quilt type (compressed/ uncompressed), hemisphere and quilt segment length (30, 60, 120, 240, 480, 960 ms) did not reveal any significant main effects or interactions (all $P > 0.1$).

To test whether the parametric response to quilts of compressed speech would plateau earlier than the uncompressed response, we fitted it with the same piecewise linear elbow function. For both the compressed and uncompressed data sets, we computed the best-fitting functions for 10,000 bootstrapped samples and computed confidence intervals on the elbow points from the resulting parameter distributions. By this analysis, the elbow point for the response to compressed quilts was significantly different from 240 ms ($P < 0.05$), and not significantly different from the elbow in the uncompressed speech quilt response ($P > 0.05$).

Taken together, these results are consistent with the hypothesis that speech analysis in the fROI is mediated by receptive fields at the scale of ~500 ms and smaller, roughly between the scale of syllables and words at normal speech rates. When speech is compressed, the overall response is lower, presumably because compressed speech deviates from the

structure of natural speech (Fig. 7a), and is therefore not an optimal stimulus for this part of the auditory system. However, compressed speech apparently contains enough natural structure that quilts from longer segments produce a higher response than quilts from shorter segments, up to the limits of the analysis window. Notably, this response increase occurred up to 480 ms, even though 480-ms segments of compressed speech are generated from 960 ms of uncompressed speech, which did not itself produce a response increment, potentially because the underlying receptive fields are not long enough to register structure over such durations.

One natural question is whether the nature of the sensitivity to speech temporal structure might vary across subregions of the fROI (for example, different subregions might exhibit elbow points at different segment lengths). To address this issue we used a parcellation algorithm[29] to identify plausible subregions in the activation locus of the functional localizer contrast, the responses of which we analyzed individually. This analysis did not reveal evidence for further differentiation: anterior and posterior parcels in temporal regions showed similar parametric responses to speech quilts (Fig. 8). Furthermore, a clustering algorithm[30] that searched for groups of voxels with similar parametric response profiles, irrespective of the voxel locations, yielded a single cluster whose mean response profile resembled the mean response of the fROI (Supplementary Fig. 4), even though in principle it could have yielded multiple clusters with different asymptotic behavior. Thus, we found no evidence for distinct forms of temporal selectivity at the voxel level, voxels were either sensitive to the quilting manipulation or not, and those that were sensitive exhibited a consistent parametric response.

DISCUSSION

We identified a bilateral region in human STS that is tuned to the acoustic structure of speech, particularly its temporal structure. This region responded more to quilts composed of long segments of speech than to quilts composed of short segments, even though these stimuli had comparable long-term and short-term spectral structure. The region is evidently sensitive to acoustic structure independent of lexical-semantic and syntactic content, as we exclusively used speech from an unfamiliar language. We found that the region's response could not be explained by amplitude modulation sensitivity or by sensitivity to prosodic pitch variation. Its sensitivity to the extent of temporal structure was apparently speech specific, in that non-speech quilts elicited low responses that did not differ substantially as a function of segment length. The region also exhibited parametric sensitivity to the degree of temporal structure: its response increased as the timescale of naturalistic speech structure increased, but only up to a point. This plateau also occurred, and at a comparable timescale, for quilts made from speech that was time-compressed. Overall, our results demonstrate a locus of speech-specific acoustic analysis and are consistent with the idea that it operates at a timescale up to that between syllables and words.

Speech specificity

The degree to which particular cortical regions are selectively involved in processing speech signals remains controversial. Both functional magnetic resonance imaging (fMRI)[31] and

electrocorticography[32] recordings have identified superior temporal lobe regions that display sensitivity to attributes of speech, typically at the level of single phonemes. However, the specificity of such regions to speech has remained unclear. For example, although domain specificity is commonly proposed for other perceptual and cognitive systems (for example, see refs. 20,33), the notion that temporal lobe regions are generic in their processing is a popular conclusion of auditory imaging reviews[34,35]. Moreover, much prior work has focused on manipulations of intelligibility that typically affect both speech acoustics and linguistic content, leaving distinctions between speech processing and linguistic processing unresolved[3,9–11]. We introduced sound quilting as a new method for manipulating temporal structure in audio signals and applied it to foreign speech, isolating responses to the acoustic-phonetic structure of speech that are distinct from those to higher order, lexically or syntactically driven structure. The observed responses to speech quilts, coupled with our control experiments with non-speech quilts, provide evidence for speech-specific processing that is distinct from linguistic processes.

In particular, our results demonstrate that speech analysis is not simply driven by amplitude modulation (AM). Syllable-rate AM is a prominent feature of speech signals and has been found to produce stronger responses in human auditory cortex than modulations at faster rates[14–16,24,25]. However, we found that control quilts whose AM characteristics were matched to those of our speech stimuli produced weak overall responses that were independent of quilt segment length, qualitatively different from the responses to speech quilts. These results do not exclude a role for slow AM in speech analysis, perhaps in the context of parsing the signal[36], but suggest that AM can only be part of the story—the STS region that we identified is apparently more specifically tuned to speech-like spectro-temporal structure. Our results also suggest some degree of tuning to speech-like spectral structure: responses to speech quilts of even the shortest segment lengths used always considerably exceeded those to each of the non-speech control stimuli.

Hierarchy of selectivity

Our results reveal clear differences between primary auditory cortex and the regions that we found to be selective to speech structure, extending existing evidence for a hierarchical organization of auditory cortex[4–10,37–39]. Whereas the STS speech region differentiated between long- and short-segment quilts, as well as between speech quilts and every other sort of quilt that we presented, HG displayed an overall response that was largely stimulus-invariant. This likely reflects the similarity of our stimuli in overall spectral and modulation power, features that are often argued to drive primary auditory regions[37]. For the most part, PT was similarly insensitive to our main stimulus manipulations. Speech-specific analysis is evidently performed subsequent to several stages of more generic auditory processing in the auditory pathway. Our results failed to reveal functional differentiation[40] in the area identified by our functional localizer (for example, in the shape of the parametric response), but are obviously limited by the spatial resolution of the hemodynamic response. It remains conceivable that subregions in the fROIs that we measured have distinct functional roles that could be revealed by additional experimental manipulations.

Lateralization

The cortical structures for processing speech have historically been attributed to the left or 'dominant' hemisphere. After decades of debate, the emerging consensus is that processing becomes progressively more left-lateralized as signals become progressively more speech-like[41]. Our results provide some support for this view, in that the effect of segment length was somewhat more pronounced in the left hemisphere fROIs than in the right. However, it is apparent from the activation maps (Fig. 2), the parametric response curves (Fig. 3) and the non-significant effect of fROI volume that the hemispheric effects are modest: the effect of speech segment length was robust in both hemispheres. It is therefore possible that laterality effects are driven more by higher order linguistic processing demands than by speech analysis *per se*[6,42,43].

Analysis timescale

Clues to the analysis occurring in the STS region we identified may be found in the response plateau at segment lengths of around 500 ms. Because the response of any speech-selective mechanism seems likely to be a highly nonlinear function of the sound waveform, we currently lack a quantitative model with which to generate precise predictions of the region's response. However, the observed response plateau is, at least naively, suggestive of neuronal mechanisms that analyze speech-related acoustic structures less than ~500 ms in duration. On the assumption that such mechanisms are tuned to the structure of natural speech, their response should increase as the temporal extent of naturalistic structure increases, but this response increase should be limited by the temporal integration limits of the underlying neuronal populations. The parametric response to quilts of time-compressed speech also exhibited a plateau at ~500 ms, indicating that the response pattern is not simply determined by the intrinsic temporal dependencies of the stimulus and could reflect an architectural property of that part of auditory cortex. One possibility is that the STS is involved in analyzing acoustic structures in speech ranging from single phonemes (20–80 ms) to pairs of syllables (250–500 ms). These acoustic-phonetic primitives could then be passed on to linguistic mechanisms that would match their sequential structure to lexical representations[2,4].

It is noteworthy that speech contains acoustic structure at time-scales longer than 500 ms. Indeed, the results shown in Figure 7a indicate that these dependencies exist and that human listeners are sensitive to them, as perceived naturalness did not plateau at 500 ms. One possibility is that most dependencies beyond 500 ms are a result of relatively generic fluctuations in pitch and amplitude that are found in many natural sounds[44,45], and as such are extracted elsewhere[46].

Imaging and recent electrophysiological data[32,47] suggest that cortical fields in lateral STG mediate the processing of subphonemic features, whereas lexical-level processing requires the middle temporal gyrus (MTG)[48,49]. However, the features of speech that comprise phonemes are typically short[17,18]. Our data raise the possibility of an intermediate processing stage in STS in which features and phonemes are integrated over time to extract longer-scale speech structures and build syllabic or lexical-size perceptual representations. Such representations could trigger or facilitate the linguistic processing that underlies

comprehension. A potential STG-STS-MTG processing hierarchy makes reasonable functional anatomic sense, but obviously merits follow-up.

Open questions

Our results suggest a number of potentially important future directions. Analogous quilting experiments using speech in a familiar language could conceivably produce stronger lateralization[41] and could shed light on the relation between speech-specific temporal analysis, intelligibility and linguistic processing. Given that German and English have considerable phonological overlap, it would also be informative to examine segment-length effects in languages that are more foreign in their phonemic inventory and phonological structure (for example, Mandarin Chinese or Zulu for English listeners). It will also be important to clarify the relation between our effects and putative voice-sensitive regions[39]; speech quilts arguably sound increasingly like realistic voices as the segment length increases. Much could also be learned from responses to speech quilts in non-human primates. Potentially homologous voice-selective regions have been identified in macaques[50], but the nature of their temporal selectivity remains unclear, as does their relation to speech-selective brain regions in humans. Measuring responses in non-human animals to quilts made from speech and from species-specific vocalizations could provide insight into the evolutionary origins of human speech analysis.

ONLINE METHODS

Participants

17 unique participants (mean age = 22.45, range = 18–30) took part in the fMRI experiments (Supplementary table 1). 16 additional participants (mean age = 35.8, range = 21–65, 9 females) took part in a behavioral experiment rating sound quilt naturalness. All participants were native speakers of American English, with no knowledge of German. All participants provided informed consent in accordance with the New York University Committee on Activities involving Human Subjects and the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects. One participant (KD, see Supplementary table 1) was excluded from further analysis because he failed to show a reliable response for the functional localizer contrast in both hemispheres, and thus we could not define fROIs. One participant (DC, see Supplementary table 1) only contributed data for the compressed speech stimulus set (S50) since the data from the repeat session (S51) were unrecoverable because of a trigger malfunction. No statistical methods were used to predetermine sample sizes but our sample sizes are similar to those reported in previous studies[51]. Participants were semi-randomly assigned for their initial scanning session; subsequent sessions were constrained such that no participant took part in the same control experiment twice.

Quilting algorithm

The quilting algorithm (Fig. 1a) generated sound signals by rearranging segments of a source signal. The goal of the algorithm was to preserve stimulus properties at a specified short timescale as best possible but to leave structure at longer timescales otherwise unconstrained. This was achieved by ordering segments such that adjacent segment pairings

were distinct from those in the source signal but such that the magnitude of the change across segment borders (in a simulated cochleogram) was otherwise as close as possible in magnitude to that in the source. Segments were then concatenated using pitch-synchronous overlap-add (PSOLA)[52] to eliminate phase discontinuities.

The algorithm included the following steps: (1) Choose a segment duration, and divide the source signal into segments of that length. (2) Assign a randomly chosen segment from the source signal (uniformly distributed over the length of the source) to be the first segment of the stimulus. We denote the position of the chosen segment in the source signal with the index *k*. (3) Compute the average change in the cochleogram (using an L2 metric) between the right-hand border of the chosen segment and the left-hand border of the remaining segments in the source signal, where the border is taken to be 30 ms in length:

$$d(k,n)=\sum_{t,f}\left[C_k^R(t,f)-C_n^L(t,f)\right]^2$$

C_k^R and C_n^L denote the right and left borders of cochleogram segments *k* and *n*, respectively.

$$C_j^R=C(j*w-30{:}j*w,1{:}F)$$

$$C_j^L=C((j-1)*w+1{:}(j-1)*w+30,1{:}F)$$

Here the cochleogram *C* is indexed in ms for simplicity, *w* is the segment length in ms and *F* is the index of the maximal frequency channel. (4) Choose the next stimulus segment to be that which gives a segment-to-segment change that is closest to the segment-to-segment change between segment *k* and segment *k* + 1 in the source signal.

$$i=\mathrm{argmin}_n|d(k,n)-d(k,k+1)|,n\neq k+1,n\notin S$$

where *S* is the set of all previously used segments. (5) Repeat steps 1–3 for the segment just chosen until the segment re-ordering is of the desired length. (6) Concatenate segments via PSOLA by (a) shifting the boundaries of segment *n* forward or backwards by at most 15 ms to maximize the cross-correlation between the waveforms of segment n and of the preceding stimulus, over the 30-ms region straddling the transition between segments; (b) cross-fading between segment n and the preceding stimulus using 30-ms raised cosine ramps centered on the segment boundary (such that a 15-ms buffer region is included on each end of each segment).

Cochleograms were computed using the auditory filterbank from ref. 26 (30 filters spanning 20 to 10,000 Hz, equally spaced on an ERB_N scale), by raising the Hilbert envelopes of the resulting subbands to a power of 0.3 (to simulate cochlear compression). Matlab code for implementing the quilting algorithm is available at http://mcdermottlab.mit.edu/.

Segment sizes were log-spaced multiples of 30 ms: 30, 60, 120, 240, 480 and 960 ms.

Stimuli

Quilts were generated from 20-s source signals, all of which were resampled to 20 kHz and bandpass filtered between 80 and 8,500 Hz before quilting (third order Butterworth filter, forward and reversed filtered to avoid phase shifts). All stimuli were 6 s in length. A linear fade was applied to the last second of each stimulus.

Speech source signals

Speech source signals were excerpted from recordings made by eight German-speaking volunteers reading from a German book. Breaths and pauses were excised from the recordings, producing ~12 min of continuous speech. The resulting speech source signals had syllable rates in the normal range (Supplementary table 2).

Modulation and co-modulation control stimuli

Modulation control source signals were generated with the texture synthesis algorithm of McDermott and Simoncelli[26]. Synthetic signals (20 s in length) were generated that matched either (1) the envelope marginal statistics and modulation power or (2) the envelope marginal statistics, modulation power and cochlear correlations of each of the 20-s speech source signals used for speech quilt generation. Statistics were measured and imposed using the auditory model specified in the original paper[26] (code for the texture synthesis is available at http://mcdermottlab.mit.edu/). These stimuli control for many standard acoustic properties but do not replicate all aspects of the temporal structure of the source signal, including dependencies between different modulation frequencies in a subband envelope, or correlations between the envelopes of different subbands at different points in time (as produced by formant transitions, for instance). Modulation spectra in Figure 4a were generated by measuring the power in a set of logarithmically spaced modulation filters[26] (each of which was applied to the envelope of a cochlear filter) for each stimulus, averaging across stimuli in each condition, and then dividing each power value by the total average power across all filters for that condition. Correlation matrices in Figure 4c were generated by averaging the correlation matrices over each stimulus in a condition (the correlation matrices were computed from the envelopes of a set of simulated cochlear filters[26]).

Environmental sounds

Environmental sound source signals were selected to have extended temporal structure. A large set of environmental sounds was initially used to generate quilts, and 18 sounds for which quilting produced the most salient effect (as subjectively judged by the authors) were chosen for the experiment. These sounds comprised (1) cars racing around a track, (2) fireworks, (3) firecrackers, (4) a pile driver, (5) raking leaves, (6) brushing teeth, (7) a person running up stairs, (8) a person jogging on gravel, (9) a person walking on gravel, (10) people marching, (11) ping pong, (12) sawing wood with a handsaw, (13) sanding wood by hand, (14) British church bells, (15) country church bells, (16) dogs barking, (17) pigs squealing and grunting, and (18) a bird singing. All of these source signals had epochs of silence removed using the same procedure used for the speech recordings.

Noise-vocoded speech

Noise-vocoded speech was generated by decomposing the speech source signals into frequency subbands, measuring their envelopes and then imposing these envelopes on subbands of white noise. Subbands were generated in the frequency domain with a filter bank that included 10 bandpass filters with center frequencies evenly spaced on an ERB_N scale, along with lowpass and highpass filters on the ends of the spectrum such that the summed response of the filter bank was flat over the entire spectrum. Center frequencies of the bandpass filters ranged from 120 to 7,067 Hz. Adjacent filters overlapped by 50% in the frequency domain, with half-cycle cosine frequency responses. The lower absolute cutoff of the lowest bandpass filter was 20 Hz, and the higher absolute cutoff of the highest bandpass filter was 10 kHz. Subband envelopes were measured as the magnitude of the analytic signal obtained via the Hilbert transform, downsampled to 100 Hz.

Speech envelopes were imposed in 30 iterations of the following set of steps. (1) Measure envelopes in noise-vocoded stimulus (initialized as white noise). (2) Divide noise-vocoded subbands by their envelopes to yield the subband fine structure. (3) Multiply the noise-vocoded subband fine structure by the envelope of the corresponding speech subband (upsampled to the 10 kHz sampling rate of the subband). (4) Filter the subbands with the same filters used to produce them (to enforce their bandlimits, as is standard in analysis-synthesis subband transforms[26]), and then sum the results to yield a full-bandwidth noise-vocoded stimulus.

This procedure differs from that used conventionally[27] in the use of overlapping filters and an iterative imposition procedure. The filter overlap has the advantage of generating a signal whose envelopes are similar to those of the source signal regardless of the exact filter bank used for measurement. Iteration has the advantage of ensuring that the generated signal has the desired envelopes. It is necessary because the subbands overlap, because the envelope imposition does not always respect the subband bandlimits, and because the envelope and fine structure are not independent, such that a single iteration of this procedure typically produces a signal whose envelopes deviate significantly from the target speech envelopes. As this procedure is repeated the envelopes converge to the desired values, and the fine structure of the noise-vocoded signal relaxes to a state that is consistent with the envelopes.

Time-compressed speech

Speech was time-compressed using the pitch-preserving compression/dilation algorithm in Praat.

Experimental design

All stimuli were presented using Psychophysics Toolbox Version 3 (ref. 53). Ten conditions were presented in each scanning session. Two ‘localizer’ conditions (speech quilts with 30- and 960-ms segments, respectively; L30 and L960) were always included, and were used to define fROIs. Six parametric speech quilt conditions (S30, S60, S120, S240, S480, S960) were also always included (for the compressed speech experiment these quilts were made from compressed speech), and were used to characterize the response with data independent of that used to define the fROIs. Finally, two control conditions (quilts made from various

non-speech source signals, with segment lengths of 30 and 960 ms; C30 and C960) served to investigate alternative explanations of the effects of interest.

A scanning session consisted of four 'runs' lasting 17 min each. Stimuli were presented in a pseudo-randomized fashion that boosted contrast selectivity. Each condition was presented 36 times per scanning session (in addition to 36 silent trials of 10-s duration) with a mean inter-stimulus interval of 4 s (range 3–5 s). Sound quilts in the L30 and S30, as well as the L960 and S960 conditions, respectively, used different exemplars. Participants were asked to keep their eyes open and to press a button at the end of each stimulus. Stimuli were presented at a comfortable listening level (~75 dB SPL) via Sensimetrics MRI-compatible insert earphones (Model S14); participants wore protective earmuffs to further reduce the scanner noise.

Repeat sessions

To maximize power within individual participants, each participant was scanned in multiple sessions when possible (1–4 sessions). Across repeat scanning sessions conditions 1 to 8 (that is, L30, L960, and S30 to S960) were fixed (though we used two sets of speakers across experiments to minimize speaker familiarity effects); the two control conditions varied. The exception to this was the compressed speech experiment, in which participants listened to the two functional localizer conditions (L30, L960), a set of parametrically varied quilts of temporally compressed speech (Comp30, Comp60, Comp120, Comp240, Comp480 and Comp960 quilts), and two other conditions whose stimuli turned out to have artifacts and were not analyzed. The participants in the compressed speech experiment were also scanned with analogous uncompressed speech conditions (12 of the 17 unique participants; see Supplementary table 1) to enable direct comparison.

Image acquisition

T2* gradient-weighted echo-planar images (EPI) were acquired on a 3-T Siemens Allegra system using a Nova Medical NM-011 head coil. 30 slices (2 × 2 × 2-mm voxels) were acquired for each volume (time to repeat (TR) / time to echo (TE): 2,100 / 30 ms; flip angle: 90°; field of view (FOV): 224; acquisition matrix: 112 × 112). The acquisition volume was tilted forward such that slices were parallel to and centered on the superior temporal gyrus. For each of four runs, 484 volumes were acquired (total of 1,936 volumes per session). After the second run, a calibration scan was acquired to allow correction of inhomogeneities of the B0 field in the EPI images. A structural high-resolution T1-weighted MRI (MPRAGE) scan (TR/TE: 2,500/3.93 ms; FOV: 256) was acquired for each participant.

Data analysis

Data collection and analysis were not performed blind to the conditions of the experiments. Imaging data were analyzed using Statistical Parametric Mapping software (SPM8, http://www.fil.ion.ucl.ac.uk/spm). The first four volumes in each run were discarded to control for T1 saturation effects. The remaining 1,920 scans were realigned to the first volume in the first block, un-warped to correct for motion artifacts and re-sliced using sinc interpolation (SPM8, realign and unwarp); the structural scan of each participant was coregistered to the mean functional scan (SPM8, coregister), segmented and spatially normalized to

standardized stereotaxic MNI space (SPM8, segment), before applying the resulting linear transformations to the EPIs and structural scan (SPM8, normalize: write). Finally, the EPIs were spatially smoothed to improve the signal-to-noise ratio using an isotropic 6-mm full-width at half-maximum (FWHM) Gaussian kernel.

The design matrix for each participant consisted of ten regressors (corresponding to the ten experimental conditions), derived by convolving the stimulus response function (modeled as a 6-s box-car function) with SPM's canonical hemodynamic response function. The silent periods were not modeled explicitly. Data were high-pass filtered at 1/128 Hz to remove slow drifts in the signal.

For each participant, we calculated an individual fROI by contrasting the two functional localizer conditions [L960 > L30] using a *t* test. For participants that were scanned more than once, we derived the fROI by calculating the [L960 > L30] contrast using data combined across repeat sessions. Individual fROIs (one per hemisphere, per participant) comprised voxels that (a) survived a statistical threshold ($P < 0.001$, uncorrected for multiple comparisons, for participants who were scanned up to two times, or $P < 0.05$, FWE, corrected, for participants who were scanned more than twice), and (b) lay within the superior temporal lobe.

Group-level analyses were based on a random-effects model within the context of the general linear model[54]. For group-level analyses, the smoothing of contrast images was increased to an effective 8 mm FWHM Gaussian kernel. The contrast images for the [L960 > L30] functional localizer contrast were subjected to a second-level one-sample *t* test; the group fROIs (one per hemisphere) comprised voxels that (a) survived a statistical threshold ($P < 0.0001$, uncorrected for multiple comparisons), and (b) lay within the superior temporal lobe. The group fROI was analyzed both to provide a general picture of the anatomical distribution of our effects, and because it seemed desirable to compare the (group-derived) anatomical ROIs to a group-derived functional ROI.

In addition to the two functional ROIs, we used two anatomical ROIs in HG (roughly corresponding to primary auditory cortex) and PT (part of non-primary auditory cortex), based on previously described probability maps (refs. 55 and 56, respectively). Both ROIs were thresholded such that they only included voxels with at least 30% probability of belonging to either structure.

The percent signal change in these four ROIs was calculated using MarsBaR[57]. Because overall response levels varied across participants and ROIs, we normalized the percentage signal change for each condition and ROI by the ROI response to the L960 condition.

For the parcellation algorithm[29], the four runs of each participant's experimental session were used to obtain functional ROIs for the Localizer contrast ([L960 – L30]); as above, repeat participant's runs were concatenated so that the parcellation algorithm worked on 4 runs for each unique participant. fROI parcels were based on a statistical threshold of $P < 0.001$ (uncorrected), an overlap threshold across participants of 1/(number of unique participants – 1), an overlap threshold for ROIs of 0.5, and 8-mm smoothing. This was computed for all participants and their experimental sessions, since all participants listened

to the L30 and L960 conditions. Responses to the six parametric segment length conditions (S30 to S960) were measured in the resulting five parcels in the subset of participants who were presented with them (Supplementary table 1).

For the mixture model clustering algorithm[30], all uncompressed speech quilt data from a participant were averaged to yield a single data set per subject (the average response of each voxel to each stimulus condition). These data sets were pooled together to form the input to the clustering algorithm (restricting the analysis to voxels that lay within the superior temporal lobe, as for all analyses in this paper). We searched for nine possible clusters, using 100 repetitions and mean centering to exclude cluster profiles that showed no response selectivity (that is, that were flat across segment lengths, as in HG, because the purpose of the analysis was to probe for distinct forms of sensitivity to the quilt segment length manipulation). Consistency scores were then computed for each cluster. The subsequent permutation test (which is blind to the true condition assignments) was run using 10,000 iterations and a model updating threshold of 10^{-4}.

Statistics

Data distributions were assumed to be normal, but this was not formally tested. Normalized BOLD percentage signal change data were analyzed via two-way repeated-measures ANOVAs, using the Greenhouse-Geisser correction when Mauchly's test indicated violations of the sphericity assumption. For the initial comparison (Fig. 3), factors were ROI (HG, PT, group fROI, individual fROI), hemisphere (left, right) and segment length (30, 60, 120, 240, 480, 960 ms). Planned pair-wise tests used Bonferroni correction for multiple comparisons. When comparing the responses of speech quilts with the various control conditions, we computed two-way repeated-measures ANOVAs for each ROI (HG and individual fROI) separately with factors quilt type (speech, control), hemisphere (left, right) and segment length (30, 960 ms). To evaluate lateralization, we used a paired-samples *t* test comparing the number of voxels in the individual left and right hemisphere fROIs.

Behavioral ratings (Fig. 7a) were analyzed using a repeated-measures ANOVA with factor segment length (30, 60, 120, 240, 480, 960 ms, original).

To evaluate whether the piecewise linear 'elbow' model provided a better description of the parametric response than a simpler linear model, we measured the error on left-out data. For each of the 15 participants, we computed the average parametric response across the other 14 participants, fit both functions to this average response and then measured the error on the left-out participant's data. We then performed a paired *t* test on these residuals (comparing the elbow and linear models).

For the evaluation of the response plateau (Fig. 7b), bootstrapping was performed using 10,000 samples. On each bootstrap iteration, the data was resampled by choosing random sets of either 15 or 11 subjects (for the uncompressed and compressed experiments, respectively) with replacement. For each sample, we computed the mean parametric response and fitted the elbow function to this mean response used least-squares. Confidence intervals were derived from the 2.5th and 97.5th percentiles of the resulting distribution of elbow points.

A **Supplementary Methods Checklist** is available.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank K. Doelling for assistance with data collection, G. Lewis for extensive help with visualization of the results using FreeSurfer, E. Fedorenko for assistance with the parcellation algorithm, D. Ellis for implementing the PSOLA algorithm for segment concatenation, the volunteers who kindly allowed us to record their speech, T. Schofield, N. Kanwisher and J. Golomb for helpful discussions, and N. Ding, A.-L. Giraud, E. Fedorenko, S. Norman-Haignere and J. Simon for helpful comments on earlier drafts of the manuscript. This work was supported by US National Institutes of Health grant 2R01DC05660 to D.P., a GRAMMY Foundation Research Grant to J.M.Z., and a McDonnell Scholar Award to J.H.M.

References

1. Stevens, KN. Acoustic Phonetics. MIT Press; 2000.
2. Poeppel D, Idsardi WJ, van Wassenhove V. Speech perception at the interface of neurobiology and linguistics. Phil. Trans. R. Soc. Lond. B. 2008; 363:1071–1086. [PubMed: 17890189]
3. Scott SK, Blank CC, Rosen S, Wise RJ. Identifcation of a pathway for intelligible speech in the left temporal lobe. Brain. 2000; 123:2400–2406. [PubMed: 11099443]
4. Hickok G, Poeppel D. The cortical organization of speech processing. Nat. Rev. Neurosci. 2007; 8:393–402. [PubMed: 17431404]
5. Rauschecker JP, Scott SK. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci. 2009; 12:718–724. [PubMed: 19471271]
6. Binder JR, et al. Human temporal lobe activation by speech and non-speech sounds. Cereb. Cortex. 2000; 10:512–528. [PubMed: 10847601]
7. Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA. Neural substrates of phonemic perception. Cereb. Cortex. 2005; 15:1621–1631. [PubMed: 15703256]
8. Obleser J, Zimmermann J, Van Meter J, Rauschecker JP. Multiple stages of auditory speech perception reflected in event-related fMRI. Cereb. Cortex. 2007; 17:2251–2257. [PubMed: 17150986]
9. Wild CJ, Davis MH, Johnsrude IS. Human auditory cortex is sensitive to the perceived clarity of speech. Neuroimage. 2012; 60:1490–1502. [PubMed: 22248574]
10. Giraud AL, et al. Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. Cereb. Cortex. 2004; 14:247–255. [PubMed: 14754865]
11. Obleser J, Eisner F, Kotz SA. Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. J. Neurosci. 2008; 28:8116–8123. [PubMed: 18685036]
12. Zatorre RJ, Belin P. Spectral and temporal processing in human auditory cortex. Cereb. Cortex. 2001; 11:946–953. [PubMed: 11549617]
13. Schönwiesner M, Rübsamen R, von Cramon DY. Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. Eur. J. Neurosci. 2005; 22:1521–1528. [PubMed: 16190905]
14. Boemio A, Fromm S, Braun A, Poeppel D. Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat. Neurosci. 2005; 8:389–395. [PubMed: 15723061]
15. Overath T, Kumar S, von Kriegstein K, Griffths TD. Encoding of spectral correlation over time in auditory cortex. J. Neurosci. 2008; 28:13268–13273. [PubMed: 19052218]
16. Overath T, Zhang Y, Sanes DH, Poeppel D. Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: fMRI evidence. J. Neurophysiol. 2012; 107:2042–2056. [PubMed: 22298830]

17. Greenberg, S. A multi-tier framework for understanding spoken language. In: Greenberg, S.; Ainsworth, WA., editors. Listening to Speech: An Auditory Perspective. Lawrence Erlbaum; 2006. p. 411-433.
18. Rosen S. Temporal information in speech: acoustic, auditory and linguistic aspects. Phil. Trans. R. Soc. Lond. B. 1992; 336:367–373. [PubMed: 1354376]
19. Efros AA, Leung TK. Texture synthesis by non-parametric sampling. IEEE Int. Conf. Comp. Vis. 1999:1033–1038.
20. Grill-Spector K, et al. A sequence of object-processing stages revealed by fMRI in the human occipital lobe. Hum. Brain Mapp. 1998; 6:316–328. [PubMed: 9704268]
21. Lerner Y, Honey CJ, Silbert LJ, Hasson U. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci. 2011; 31:2906–2915. [PubMed: 21414912]
22. Pallier C, Devauchelle A-D, Dehaene S. Cortical representation of the constituent structure of sentences. Proc. Natl. Acad. Sci. USA. 2011; 108:2522–2527. [PubMed: 21224415]
23. Abrams DA, et al. Decoding temporal structure in music and speech relies on shared brain resources but elicits different fine-scale spatial patterns. Cereb. Cortex. 2011; 21:1507–1518. [PubMed: 21071617]
24. Giraud AL, et al. Representation of the temporal envelope of sounds in the human brain. J. Neurophysiol. 2000; 84:1588–1598. [PubMed: 10980029]
25. Harms MP, Guinan JJ, Sigalovsky IS, Melcher JR. Short-term sound temporal envelope characteristics determine multisecond time patterns of activity in human auditory cortex as shown by fMRI. J. Neurophysiol. 2005; 93:210–222. [PubMed: 15306629]
26. McDermott JH, Simoncelli EP. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron. 2011; 71:926–940. [PubMed: 21903084]
27. Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. Science. 1995; 270:303–304. [PubMed: 7569981]
28. Davis MH, Johnsrude I. Hierarchical processing in spoken language comprehension. J. Neurosci. 2003; 23:3423–3431. [PubMed: 12716950]
29. Fedorenko E, Hsieh PJ, Nieto-Castanon A, Whitfeld-Gabrieli S, Kanwisher N. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J. Neurophysiol. 2010; 104:1177–1194. [PubMed: 20410363]
30. Lashkari D, Vul E, Kanwisher NG, Golland P. Discovering structure in the space of fMRI selectivity profiles. Neuroimage. 2010; 50:1085–1098. [PubMed: 20053382]
31. Formisano E, De Martino F, Bonte M, Goebel R. “Who” is saying “what”? Brain-based decoding of human voice and speech. Science. 2008; 322:970–973. [PubMed: 18988858]
32. Mesgarani N, Cheung C, Johnson K, Chang EF. Phonetic feature encoding in human superior temporal gyrus. Science. 2014; 343:1006–1010. [PubMed: 24482117]
33. Kanwisher N, McDermott J, Chun MM. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 1997; 17:4302–4311. [PubMed: 9151747]
34. Price C, Thierry G, Griffths T. Speech-specific auditory processing: where is it? Trends Cogn. Sci. 2005; 9:271–276. [PubMed: 15925805]
35. Schirmer A, Fox MP, Grandjean D. On the spatial organization of sound processing in the human temporal lobe: a meta-analysis. Neuroimage. 2012; 63:137–147. [PubMed: 22732561]
36. Ghitza O. On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. Front. Psychol. 2012; 3:238. [PubMed: 22811672]
37. Rauschecker JP. Cortical processing of complex sounds. Curr. Opin. Neurobiol. 1998; 8:516–521. [PubMed: 9751652]
38. Norman-Haignere S, Kanwisher N, McDermott JH. Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. J. Neurosci. 2013; 33:19451–19469. [PubMed: 24336712]
39. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. Nature. 2000; 403:309. [PubMed: 10659849]

40. Liebenthal E, Desai RH, Humphries C, Sabri M, Desai A. The functional organization of the left STS: a large scale meta-analysis of PET and fMRI studies of healthy adults. Front. Neurosci. 2014; 8:289. [PubMed: 25309312]
41. Peelle JE. The hemispheric lateralization of speech processing depends on what "speech" is: a hierarchical perspective. Front. Hum. Neurosci. 2012; 6:309. [PubMed: 23162455]
42. Cogan GB, et al. Sensory-motor transformations for speech occur bilaterally. Nature. 2014; 507:94–98. [PubMed: 24429520]
43. McGettigan C, et al. An application of univariate and multivariate approaches in FMRI to quantifying the hemispheric lateralization of acoustic and linguistic processes. J. Cogn. Neurosci. 2012; 24:636–652. [PubMed: 22066589]
44. Voss RF, Clarke J. 1/f noise in music and speech. Nature. 1975; 258:317–318.
45. Attias, H.; Schreiner, CE. Temporal low-order statistics of natural sounds. In: Mozer, MC.; Jordan, MJ.; Petsche, T., editors. Advances in Neural Information Processing Systems. Vol. 9. MIT Press; 1997. p. 27-33.
46. Meyer M, Alter K, Friederici AD, Lohmann G, von Cramon DY. fMRI reveals brain regions mediating slow prosodic modulations in spoken sentences. Hum. Brain Mapp. 2002; 17:73–88. [PubMed: 12353242]
47. Humphries C, Sabri M, Lewis K, Liebenthal E. Hierarchical organization of speech perception in human auditory cortex. Front. Neurosci. 2014; 8:406. [PubMed: 25565939]
48. Turken AU, Dronkers NF. The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. Front. Syst. Neurosci. 2011; 5:1. [PubMed: 21347218]
49. Lau EF, Phillips C, Poeppel D. A cortical network for semantics: (de)constructing the N400. Nat. Rev. Neurosci. 2008; 9:920–933. [PubMed: 19020511]
50. Petkov CI, Logothetis N, Obleser J. Where are the human speech and voice regions, and do other animals have anything like them? Neuroscientist. 2009; 15:419–429. [PubMed: 19516047]
51. Desmond JE, Glover GH. Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. J. Neurosci. Methods. 2002; 118:115–128. [PubMed: 12204303]
52. Moulines E, Charpentier F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 1990; 9:453–467.
53. Brainard DH. The psychophysics toolbox. Spat. Vis. 1997; 10:433–436. [PubMed: 9176952]
54. Friston KJ, et al. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 1995; 2:189–210.
55. Rademacher J, et al. Probabilistic mapping and volume measurement of human primary auditory cortex. Neuroimage. 2001; 13:669–683. [PubMed: 11305896]
56. Westbury CF, Zatorre RJ, Evans AC. Quantifying variability in the planum temporale: a probability map. Cereb. Cortex. 1999; 9:392–405. [PubMed: 10426418]
57. Brett M, Anton J-L, Valabregue R, Poline J-B. Region of interest analysis using an SPM toolbox (abstract). Neuroimage. 2002; 16(suppl. 2)

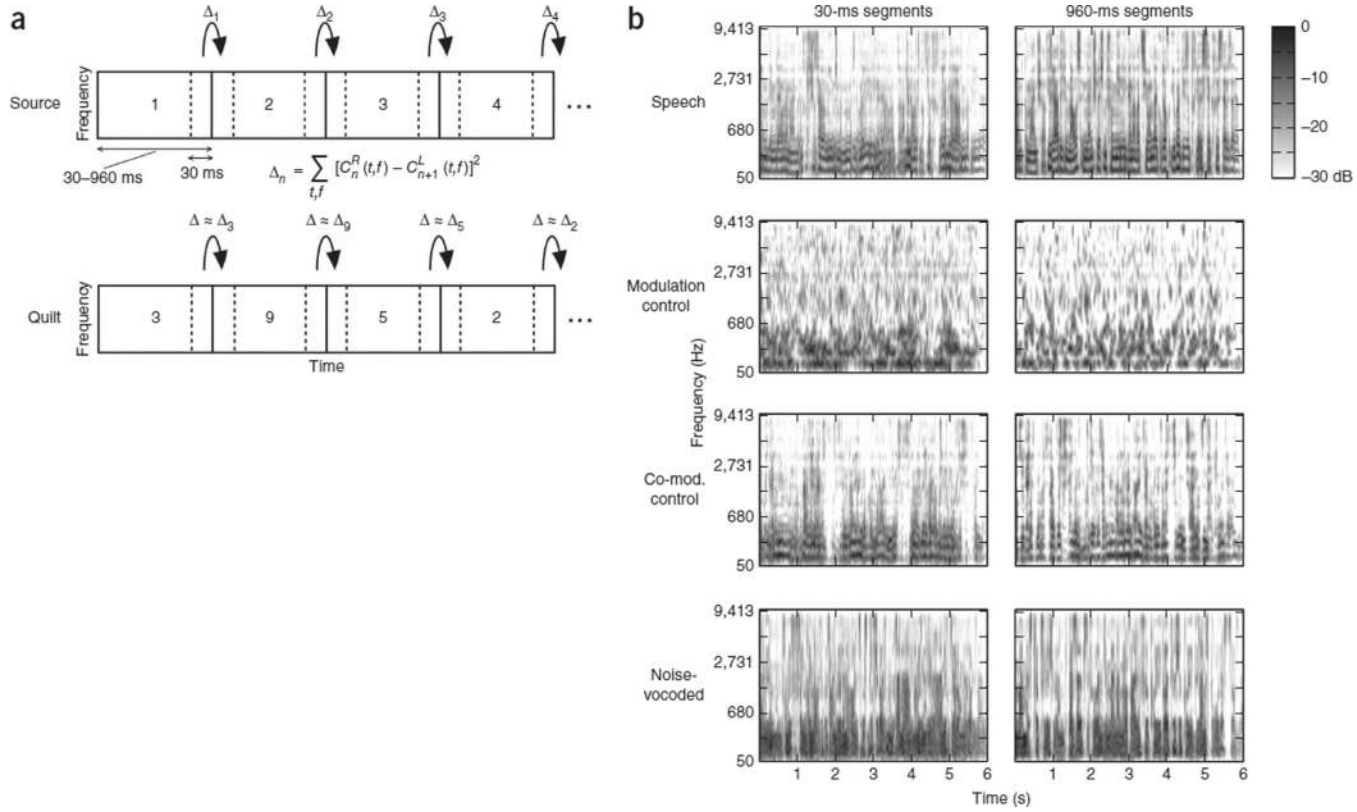

Figure 1.

Schematic of the quilting algorithm and example stimuli. (**a**) Quilting algorithm. A source signal is divided into equal-length segments (ranging from 30 to 960 ms). Segments are then reordered subject only to the constraint that they best match the segment-to-segment changes in the cochleogram of the source signal. Segment-to-segment changes were calculated from the 30-ms sections at the borders of each pair of segments, indicated by the dashed lines. In the equation defining the segment-to-segment change, $C_n^R(t, f)$ and $C_n^L(t, f)$ denote the cochleogram value at time *t* and frequency *f* of the right and the left border of the *n*th segment, respectively. (**b**) Example cochleograms of quilts made from 30- and 960-ms segments, from each of four source signals: German speech, a modulation-matched control signal, a co-modulation–matched control signal and noise-vocoded German speech. Quilts of long and short segments were not markedly different in visual appearance, but sound notably distinct in all cases.

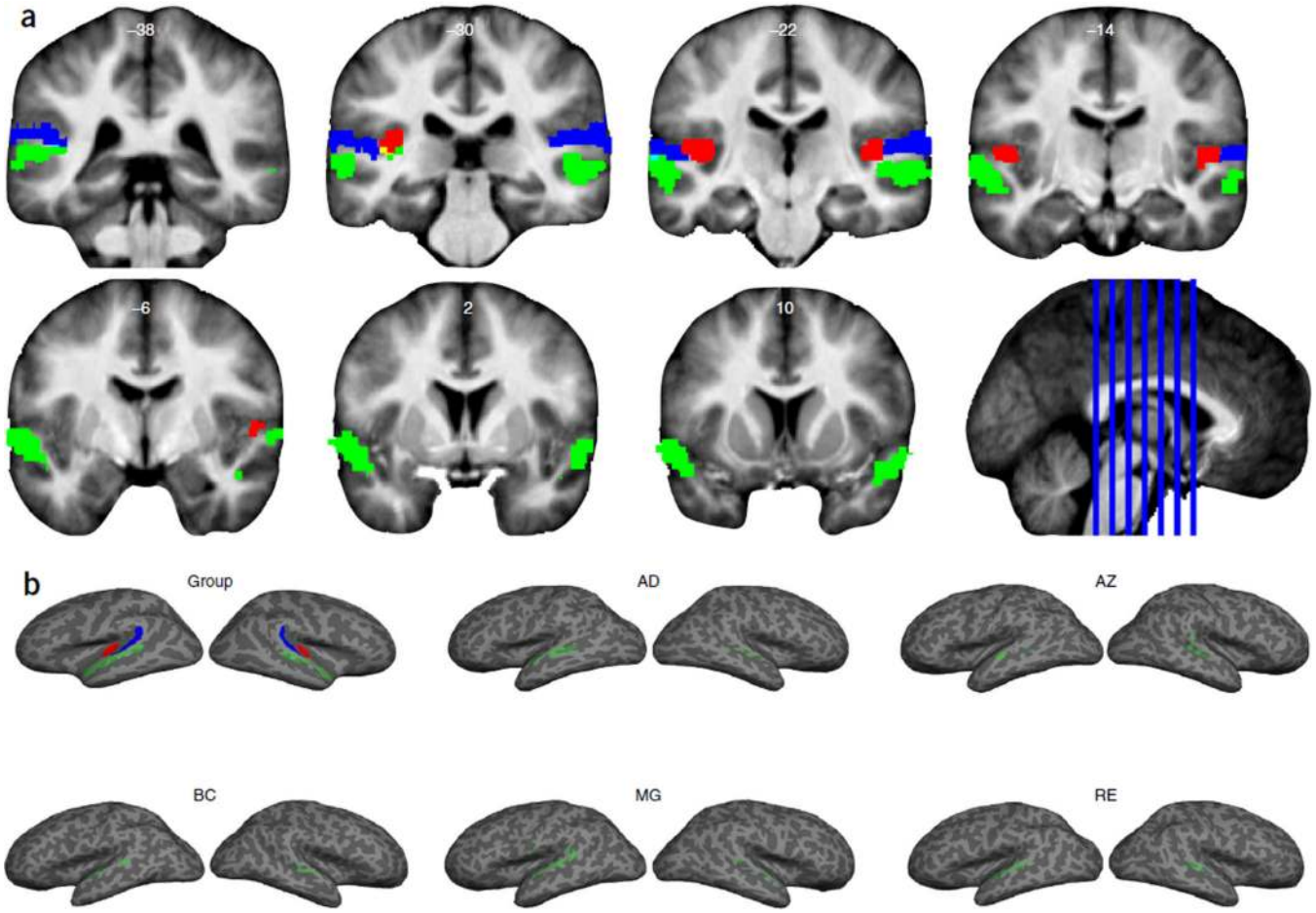

Figure 2.

Extent and location of ROIs. (**a**) Anatomical (HG and PT, red and blue, respectively) and functional (green) group ROIs displayed on coronal cross-sections of our participants' average structural images (y = −38, −30, −22, −14, −6, 2, 10). The functional group ROI was derived from the functional localizer contrast [L960 > L30], $P < 0.0001$, uncorrected. (**b**) Renderings on flattened surfaces for the three group ROIs from **a** (top left) and for individual functional ROIs for five participants who were scanned four times (rendered on their flattened structural images), $P < 0.05$, family-wise error (FWE) corrected.

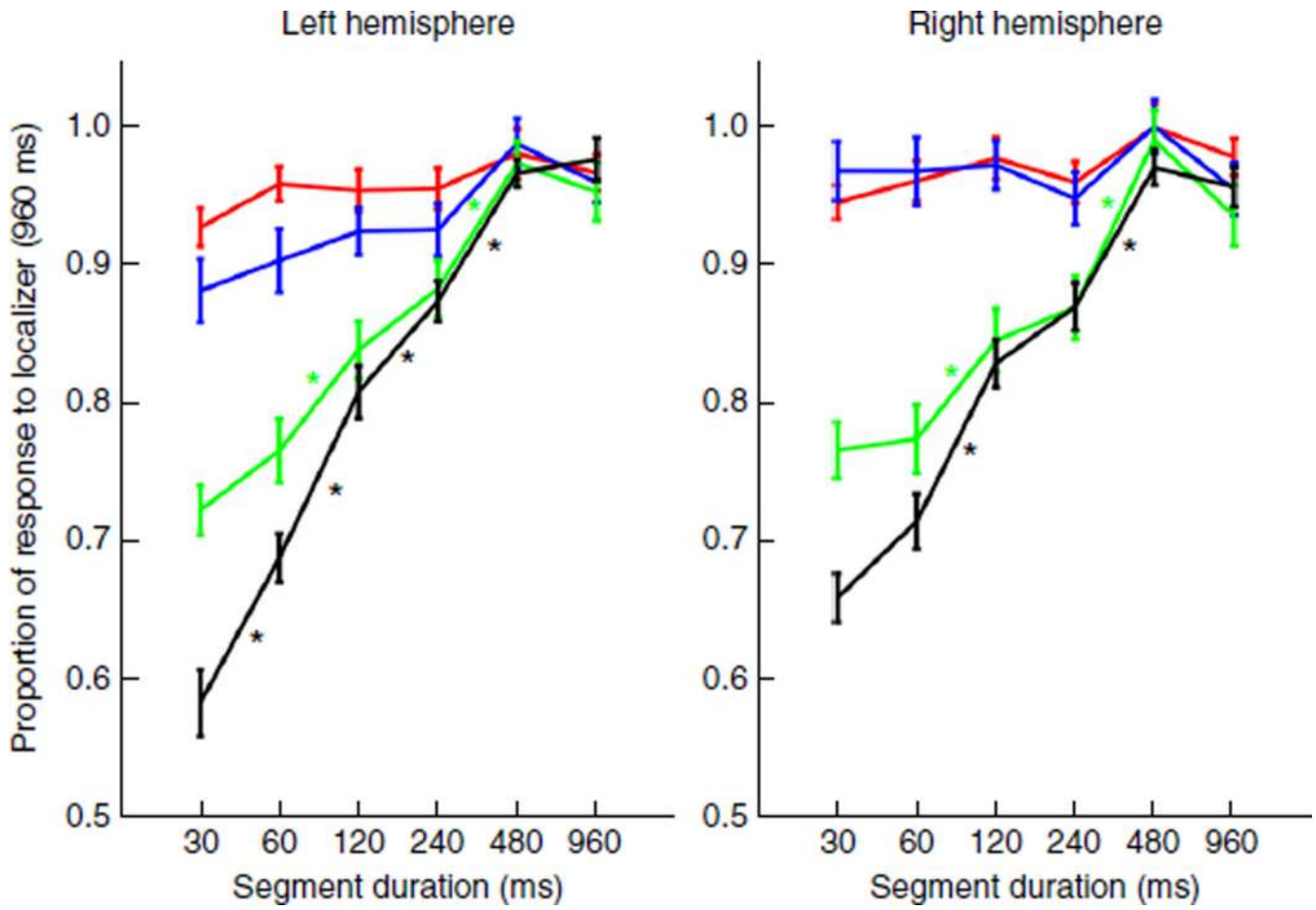

Figure 3.

Responses to German speech quilts as a function of segment length, in four ROIs: HG (red), PT (blue), group fROI (green) and individual fROIs (black), shown separately for the two hemispheres. Data are averaged across 15 unique participants. Error bars denote ±1 s.e.m., asterisks denote significant pair-wise comparisons (after Bonferroni correction), $P < 0.05$. Responses were normalized in each ROI to the response of the independent functional localizer condition L960.

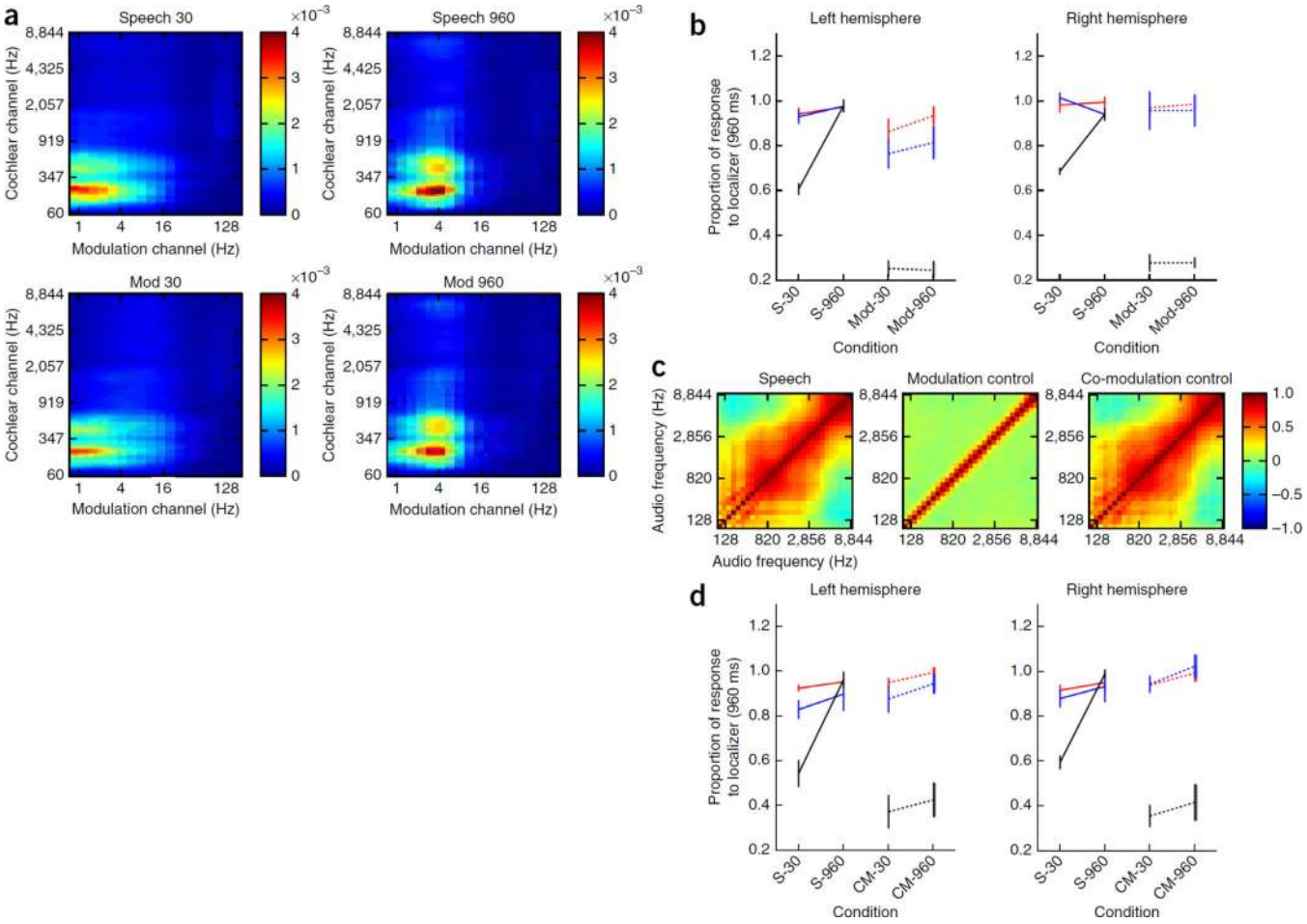

Figure 4.

Responses to modulation control stimuli. (**a**) Power in a set of simulated modulation filters[26] (normalized by the total power over all filters) for speech quilts and modulation control quilts. Note the differences across segment lengths and the similarity across quilt type. (**b**) Average responses (±s.e.m.) in HG (red), PT (blue) and the individual fROI (black) to speech quilts (solid) and control quilts (dashed) with segment durations of 30 and 960 ms. Data were averaged across the nine participants who were scanned with the modulation-control condition set. (**c**) Cross-channel correlations[26] for speech, modulation control and co-modulation control quilts (measured from 960-ms quilts). (**d**) Average responses (±s.e.m.) in HG (red), PT (blue) and the individual fROI (black) to speech quilts (solid) and co-modulation control quilts (dashed) with segment durations of 30 and 960 ms. Data are averaged across the five participants who were scanned with the co-modulation–control condition set.

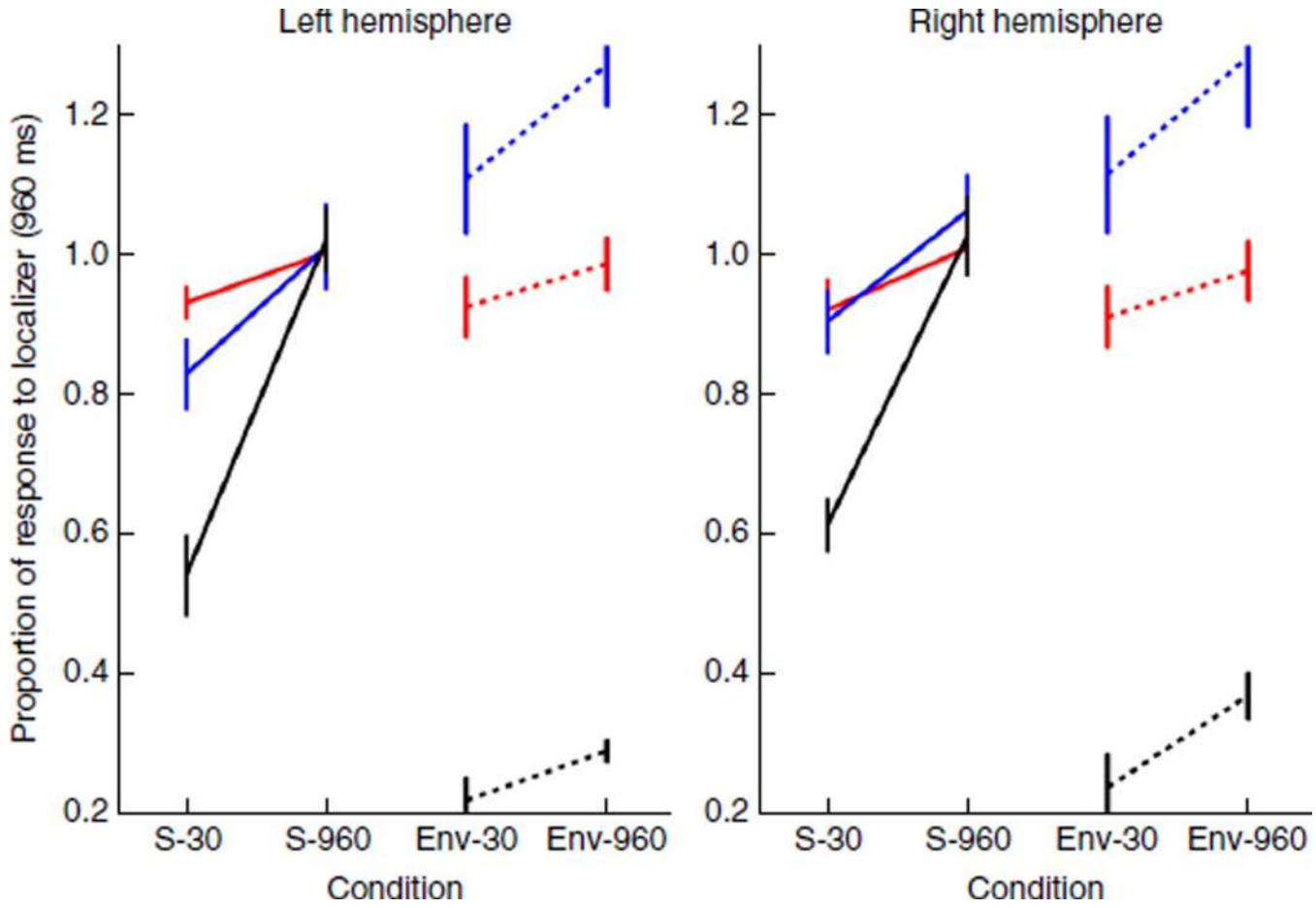

Figure 5.

Responses to environmental sound quilts. Average responses (±s.e.m.) in HG (red), PT (blue) and the individual fROI (black) to speech quilts (solid) and environmental sound quilts (dashed) with segment durations of 30 and 960 ms. Data are averaged across the five participants who were scanned with the environmental sound control condition set.

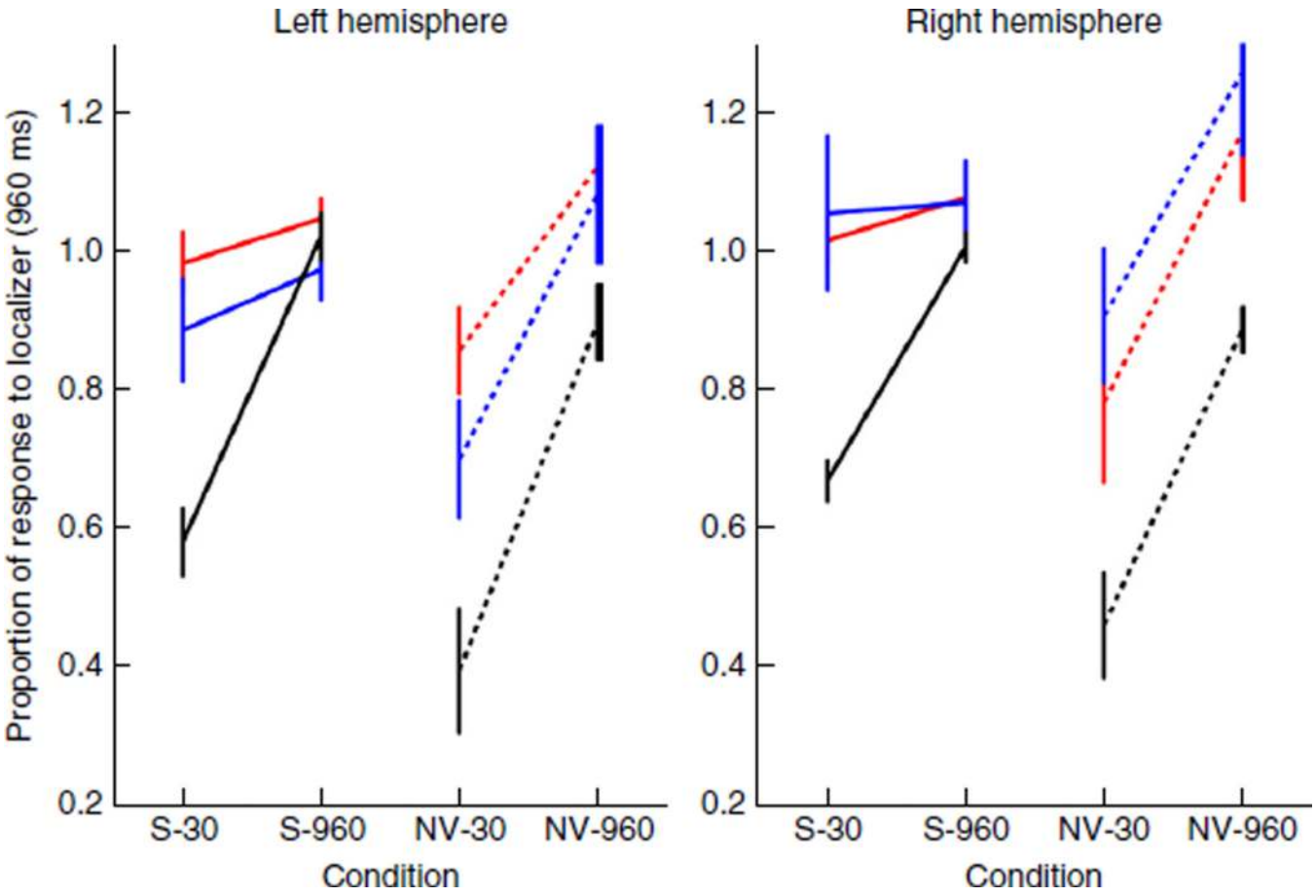

Figure 6.

Responses to noise-vocoded speech quilts. Average responses (±s.e.m.) in HG (red), PT (blue) and the individual fROI (black) to speech quilts (solid) and noise-vocoded quilts (dashed) with segment durations of 30 and 960 ms. Data are averaged across the five participants who were scanned with the noise-vocoded control condition set.

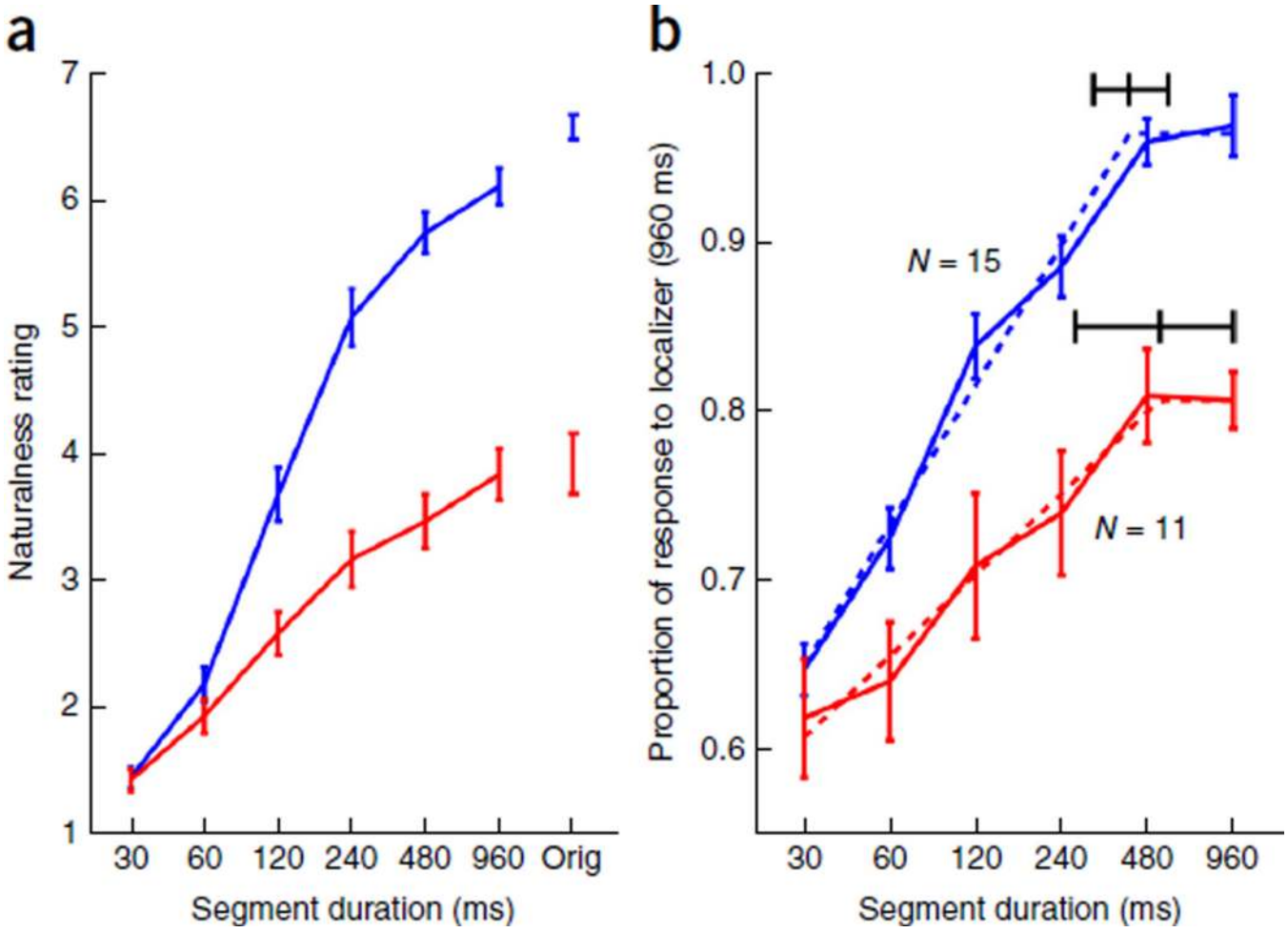

Figure 7.

Naturalness ratings and responses to compressed speech quilts. (**a**) Average ratings of the naturalness of quilted and unquilted speech stimuli for quilts made from uncompressed (blue) and compressed (red) speech (16 participants). Quilts from both quilt types were intermixed in a single block. Rated naturalness (±s.e.m.) increased monotonically with segment length for both quilt types, but the compressed speech quilts were overall rated as less natural. Naturalness ratings did not plateau at 480 ms for either quilt type. 'Orig' denotes excerpts of unquilted speech, which were included in the behavioral experiment for completeness. (**b**) Comparison of responses to quilts made from uncompressed and compressed speech. Solid lines plot average responses (±s.e.m.) to quilts of different segment lengths generated from either uncompressed (blue) or compressed (red) speech. Dashed lines plot piecewise linear function fits to the BOLD response. Black lines denote median and 95% confidence intervals on the elbow points of the fit functions for compressed and uncompressed speech, derived from bootstrap.

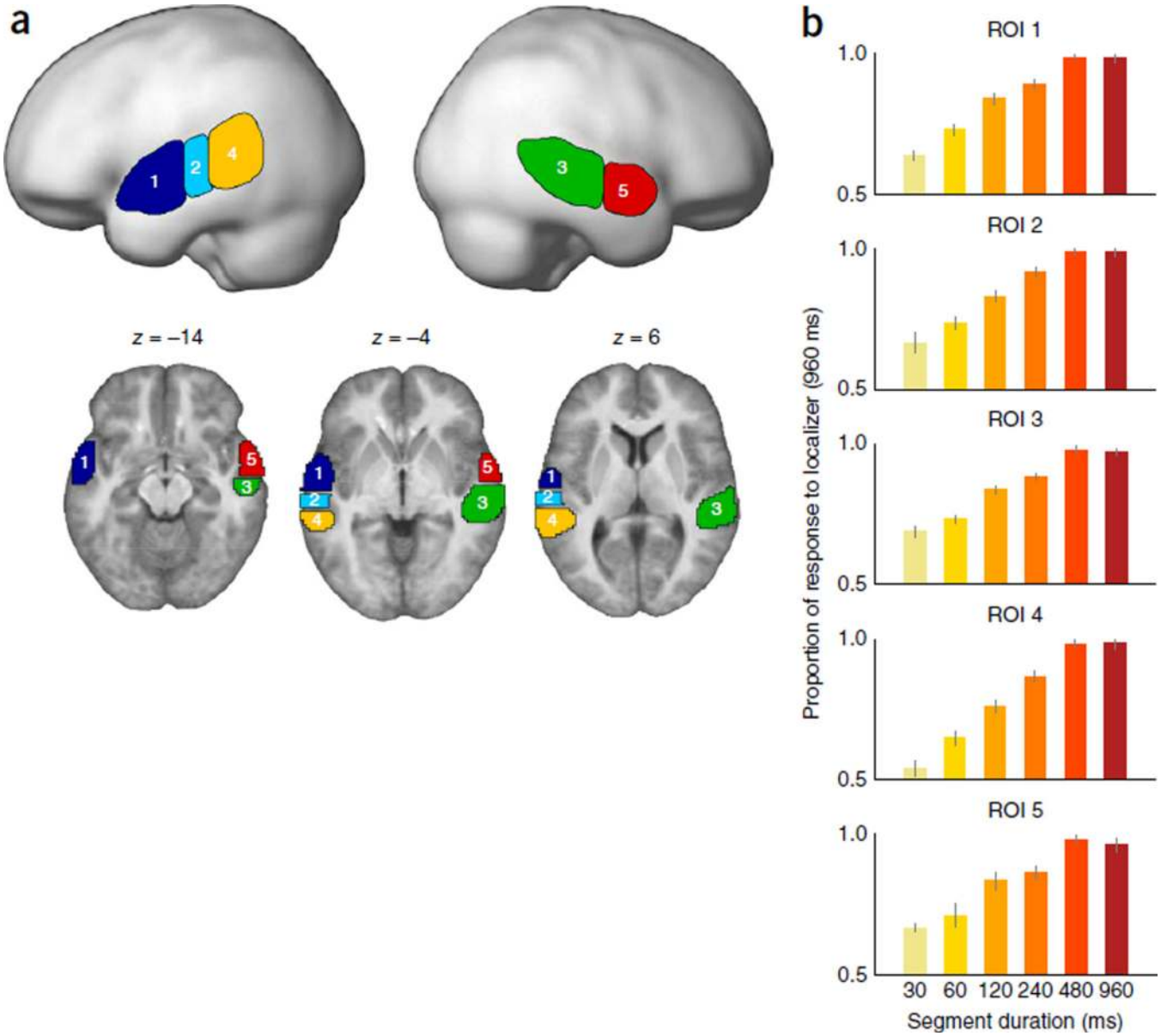

Figure 8.

Functional ROIs revealed by a parcellation algorithm[29]. (**a**) The five color-coded fROIs were rendered onto SPM's smoothed surface template (top) and on individual axial slices (bottom) of our participants' average structural images. Mean [*x*, *y*, *z*] Montreal Neurological Institute (MNI) voxel coordinates for each parcel were: [−58, −3, −6] (parcel 1); [−61, −22, 0] (parcel 2); [56, −26, −1] (parcel 3); [−59, −37, 6] (parcel 4); [59, 1, −10] (parcel 5). (**b**) The average response (±s.e.m.) to the six different segment length conditions (normalized with respect to the L960 localizer condition in a parcel) was plotted for each of the five fROIs ($n = 15$). The response pattern was similar across fROIs.